Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions

Vincent Aubanel, Martin Cooke, Chris Davis, Jeesun Kim

Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions

Vincent Aubanel
University of Grenoble Alpes, Centre National de la Recherche Scientifique, Laboratoire Grenoble Images Parole Signal Automatique, 38402, Saint Martin d'Hères Cedex, France
vincent.aubanel@gipsa-lab.fr

Martin Cooke
Language and Speech Laboratory, Universidad del País Vasco, Vitoria 01006, Spain
m.cooke@ikerbasque.org

Chris Davis and Jeesun Kim
The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith, New South Wales, 2751, Australia
chris.davis@westernsydney.edu.au, j.kim@westernsydney.edu.au

Abstract: Cochlea-scaled entropy (CSE) was proposed as a signal-based metric for automatic detection of speech regions most important for intelligibility, but its proposed superiority over traditional linguistic and psychoacoustical characterisations was not subsequently confirmed. This paper shows that the CSE concept is closely related to intensity and as such captures similar speech regions. However, a slight but significant advantage of a CSE over an intensity-based characterisation was observed, associated with a time difference between the two metrics, suggesting that the CSE index may capture dynamical properties of the speech signal crucial for intelligibility.

[RS]
Date Received: March 30, 2018 **Date Accepted:** May 22, 2018

1. Introduction

The cochlea-scaled entropy (CSE) metric (Stilp and Kluender, 2010) has been proposed as a signal-based descriptor of intelligibility. Inspired by the property of sensory systems which respond primarily to change, the authors hypothesised that the speech regions most important for recognition by human listeners can be characterised by a measure of change over time in the auditory channel. For this purpose, they defined entropy as the running average of successive differences of adjacent spectral slices, previously transformed to approximate the output of the peripheral auditory system. Using a noise replacement paradigm, they found that the intelligibility of American English sentences was linearly correlated with the proportion of entropy replaced, a pattern that was clearer than when replacing traditionally posited linguistic categories such as consonants, vowels or transitions between consonants and vowels.

While the correlation between CSE and intelligibility has been extended to Mandarin Chinese (Jiang *et al.*, 2012), subsequent studies have failed to confirm a specific role for CSE in accounting for intelligibility above and beyond a simple characterisation of the sound intensity level (Chen and Loizou, 2012; Shu *et al.*, 2016). Spectral change was posited by the original authors as the core cue for explaining their findings, but it appears that the interpretation of spectral change—and the effectiveness of the metric for predicting intelligibility—is conditioned by specific implementation choices. That is, in Stilp and Kluender (2010), the CSE is calculated based on a linear measure of magnitude differences between successive time slices; this contrasts with the classical logarithmic decibel-scaling generally used in sound perception representations and sensory systems. Indeed, as pointed out by Oxenham *et al.* (2017), a representation based on a linear scale of intensity values at the outputs of auditory filters gives greater importance to lower frequency bands.

The current paper has two aims. The first is to confirm a close link between the CSE metric and sound intensity, highlighting the importance of the choice of scale for intensity values. We hypothesised that replacing speech segments in noise would be most disruptive (and to a similar extent) for segment selection based on either intensity or CSE using linear-scaled values, and less disruptive when segments are selected based on CSE with decibel intensity scaling. A second goal is to explore the basis for the superior intelligibility predictions of the CSE characterisation, by focussing on the

temporal relationships between the metric and the underlying components of the speech signal.

2. Methods

The current study uses a noise replacement paradigm, following Stilp and Kluender (2010), in which selected segments of the speech signal are excised and replaced with noise. By measuring keyword scores in sentences that have undergone noise replacement, an estimate of the value of the information in the segments replaced by noise can be obtained. Different metrics will typically select distinct segments for replacement, enabling a comparison of the extent to which those metrics are able to predict intelligibility.

2.1 Metrics

Three metrics were contrasted: INT, CSE, and CSE-γ. As in Stilp and Kluender (2010), in each case the metric is computed in non-overlapping 16 ms frames. The INT metric is simply the root-mean-square (RMS) of the speech amplitude. The CSE and CSE-γ metrics are calculated using a 7-frame running average of the Euclidean distance between adjacent across-frequency vectors output by two distinct auditory-inspired representations of speech. The CSE metric uses the "roex" auditory filterbank (Patterson *et al.*, 1982) while the CSE–γ metric is based instead on gammatone filters (Patterson *et al.*, 1988). In the latter case, the Hilbert envelope at the output of each gammatone filter is computed and subsequently smoothed via a leaky integrator with an 8 ms time constant, log-compressed and downsampled to 16 ms frames. In both cases, 33 auditory filters were used spanning the range 26–7743 Hz. Note that the input to the distance computation in the case of the CSE metric is linear while for the CSE–γ metric it is logarithmic. The outputs of the roex filter and the gammatone auditory filterbank are shown, along with a spectrogram, in the top three panels of Fig. 1 for an example sentence.

2.2 Stimuli

Sentence material was taken from the MAVA corpus (Aubanel *et al.*, 2017), a recording of 205 phonemically balanced Harvard sentences (Rothauser *et al.*, 1969) spoken by a native Australian talker, with manually checked annotation at the word and phoneme level. Sentences include five keywords and have a mild lexical and semantic predictability. Manipulation and playback of sentences was done at a sampling rate of 16 kHz.

Following Stilp and Kluender (2010), noise replacement segments were determined based on epochs when each metric took on high or low values, leading to a set of six conditions denoted HI-INT, LO-INT, HI-CSE, LO-CSE, HI-CSE-γ, and LO-CSE-γ. The procedure for determining the high-valued segments was based on the following iterative process. The global maximum of the metric across the sentence was chosen as the centre of the first segment selected for noise replacement. The segment (of length 7 frames)

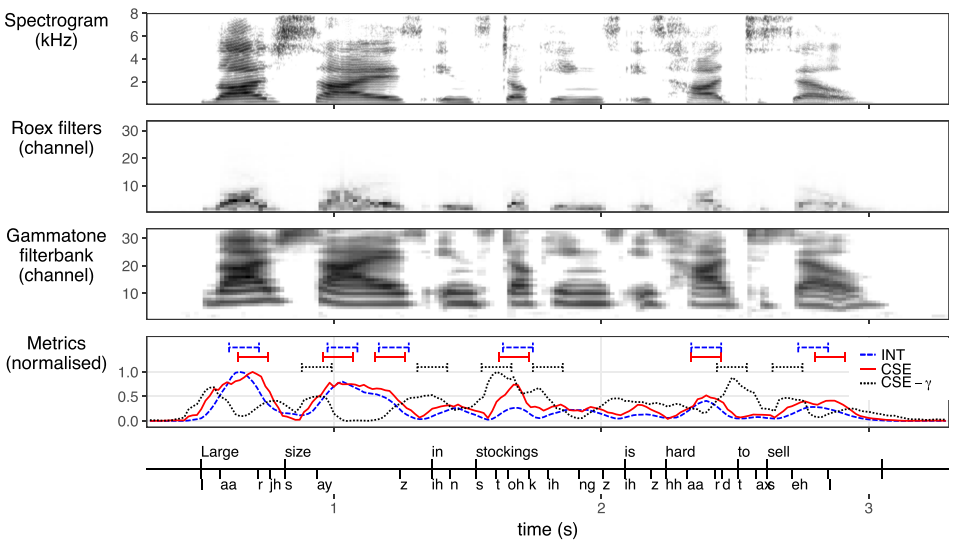

Fig. 1. (Color online) An illustration of spectro-temporal representations of the sentence "A large size in stockings is hard to sell" (top 3 panels) alongside values for the metrics tested in the current study (lower panel). The horizontal bars in the lower panel indicate the high-valued segments selected for noise replacement for each metric. Word and phoneme-level annotation is shown at the bottom.

and 5 frames on either side were then removed from consideration, and the next highest point in the metric was chosen. The procedure was repeated until no more segments could be selected. The inverse process was used to determine low value segments (i.e., choosing the smallest value of the metric at each point). The results of this procedure are visualised in the lower panel of Fig. 1, which shows segments for conditions HI-INT, HI-CSE, and HI-CSE-γ computed for the example sentence.

The above procedure typically selects 6 or more segments per utterance, but at some point the selected high-valued segments start to overlap with the low-valued segments. For the current material we determined a cross-over point at around 7 segments. To maximise the contrast amongst segments selected by the high and low variants, the procedure was truncated after the selection of 6 segments.

In addition to the experimental conditions derived from the three metrics, two control conditions were constructed. In the RANDOM condition, 6 segments meeting the above adjacency constraints were selected. In the UNPROCESSED control condition, no segments were removed (i.e., sentences were presented intact).

Selected segments were replaced by speech-shaped noise with a RMS value equal to that of the sentence after leaving out the segments to be replaced. Noise replacement was performed via an overlap-add procedure in which the speech signal was reduced to zero by 20 ms half-Hanning windows centred on segment endpoints and the masker was correspondingly amplified to unity using inverted half-Hanning windows centred on the same points. Speech-shaped noise whose long-term spectrum matched that of the talker who produced the corpus was constructed by filtering white noise with a 200-pole LPC filter computed from all sentences of the corpus.

2.3 Participants and procedure

Thirty six participants (five males) were recruited from the undergraduate student population of Western Sydney University. All participants provided informed consent, were native Australian speakers and none reported hearing problems; all received course credit for their participation. All research procedures were approved by the Human Research Ethics Committee of Western Sydney University (ref. H9495). Five participants were discarded after failing to meet a criterion level of performance (correctly identifying at least 75% of the words in the unprocessed condition). Participants were seated in sound isolated booths and listened to the stimuli through Beyerdynamic DT770 Pro closed headphones on a MacBook Pro running a custom MATLAB/PSYCHTOOLBOX script.

Participants heard a total of 176 sentences (22 sentences in each of the 8 conditions) grouped into 4 blocks of 44 sentences. After each sentence participants typed what they heard before proceeding to the next sentence. Participants could only listen to the sentence once and were able to take a short break after each block. Sentence order was randomised and condition order was balanced across participants so that for each participant, each sentence occurred only once in the experiment and each condition occurred only once in each successive group of eight sentences. Before the main test, listeners heard eight practice sentences (one for each condition) in a random order. None of the practice sentences occurred in the main experiment. The experiment took approximatively 45 min to complete.

3. Results

3.1 Listeners' responses

Figure 2 shows listeners' responses expressed as the mean proportion of keywords correct in each condition. From a baseline proportion of 0.93 correctly recognised keywords in the unprocessed condition, randomly replacing 112 ms segments by noise led to a mean score of 0.77. Both INT and CSE metrics were successful in capturing important information in speech, since selecting only low-valued regions in both of these metrics led to a mild decrease in intelligibility (0.85 in both cases) while replacing high-valued segments by noise was more disruptive (0.68 and 0.64, respectively). In contrast, selecting either high- or low-valued segments using the CSE-γ metric had no impact on listeners' scores, with an average score equivalent to a random selection (0.77 and 0.76, respectively). One notable result is that replacing high-valued segments based on the CSE metric was more damaging than equivalent segments selected by the INT or CSE-γ metrics.

A generalised linear mixed-effect model was fitted to the sentence scores [R package LME4, Bates *et al.* (2015)], predicting the probability of correctly identifying the keywords of the sentence. Experimental condition was taken as a fixed effect with 8 levels and the intercepts for participant and sentence as random effects. Random effects standard deviations for participant and sentence were 0.53 and 0.77, respectively. *Post hoc* multiple comparisons between conditions using Tukey contrasts revealed that

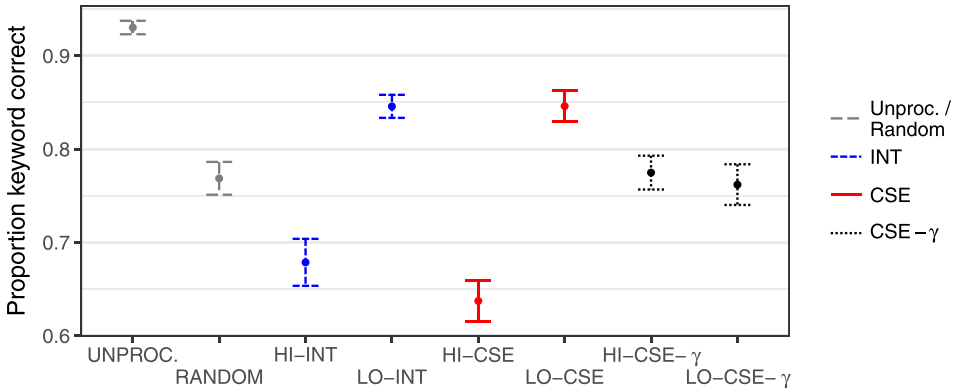

Fig. 2. (Color online) Listeners' responses expressed as the proportion of correctly recognised keywords in each condition. Error bars represent ± 1 standard error of the mean across $N = 31$ participants.

all but four pair-wise comparisons between conditions were statistically significant [all $p < 0.001$; p-values reported here and elsewhere are adjusted p-values computed with single-step method as defined in the R package MULTCOMP, Hothorn *et al.* (2008)]. In particular, word recognition in the HI-CSE condition was significantly lower than in the HI-INT condition ($z = -8.76$, $p < 0.001$). The non-significant comparisons were [RANDOM, HI-CSE-γ], [RANDOM, LO-CSE-γ], [HI-CSE-γ, LO-CSE-γ], and [LO-INT, LO-CSE] (all $p > 0.8$).

3.2 Temporal alignment of HI-CSE and HI-INT

All three metrics are based on criteria relating to the temporal dynamics of some measure of sound intensity, ultimately leading to information at discrete time instants which can be directly compared. Since HI-CSE and HI-INT segments are mainly associated with the more energetic vowel sounds of a sentence, the temporal location of segments selected by each metric in relation to vowels was examined in more detail. To evaluate the relative temporal alignment of HI-CSE and HI-INT segments, we computed the time difference between the segment center and the onset of the vowel closest to the segment center, in a window that ranged from 150 ms prior to the vowel onset to 250 ms following the onset. The upper panel of Fig. 3 plots the time to vowel onset distribution for HI-CSE and HI-INT segments, and includes the HI-CSE-γ segments for comparison.

On average, both HI-CSE and HI-INT segments occur in the portion of speech signal just following vowel onset, with HI-CSE segments located earlier than HI-INT segments. The distributions are not normally distributed (Shapiro test: CSE: $W = 0.86$, $p < 0.001$; Intensity: $W = 0.88$, $p < 0.001$) and are all leptokurtic (CSE: kurtosis = 8.1; INT: kurtosis = 10.1). A Wilcoxon rank sum test with continuity correction revealed that

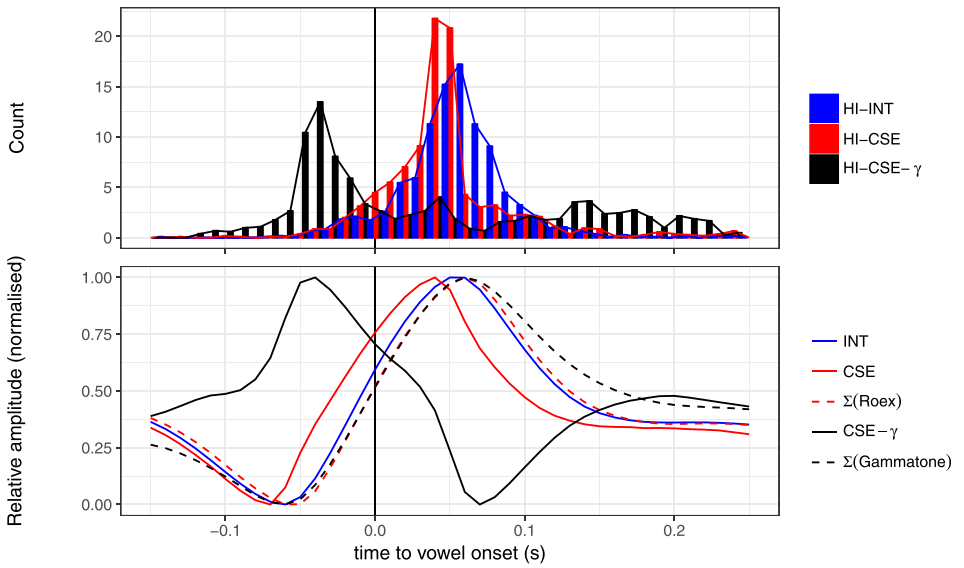

Fig. 3. (Color online) Temporal alignment of the three metrics in relation to vowel onset. Upper panel: distribution of 112 ms high-valued segments' temporal midpoints in relation to the closest vowel onset. Lower panel: temporal profile of the three metrics and of the auditory representations they are based on.

HI-CSE and HI-INT distributions have significantly different true locations ($W = 410930$, $p < 0.001$). Median values are 42.5 ms for HI-CSE and 56.4 ms for HI-INT, a difference of nearly 14 ms.

The lower panel of Fig. 3 represents the temporal profile around vowel onsets of the across-frequency sum of the auditory representations and their first order derivatives—the latter quantities being the indices CSE and CSE-γ. The profiles of the two auditory representations without the derivative step are very similar both to each other and to that for plain intensity. However, applying the smoothed first-order derivative on these auditory representations as in the CSE and CSE-γ metrics results in a marked difference. In particular, the CSE-γ profile is very different from the sum across gammatone filter envelopes on which it is based, having a form typical of a first-order derivative. In contrast, the CSE profile is quite similar to its roex base, the key difference being in the timing of the local extrema, which occur earlier.

The original study by Stilp and Kluender (2010) used both 112 and 80 ms analysis windows representing typical vowel-like and consonant-like segment durations in American English, respectively. While participants of the current study were not tested using the latter condition, the above analysis with an 80 ms windows was repeated, revealing that the relative properties of the distributions remained identical: the HI-CSE distribution had a negative mode, followed by HI-CSE and HI-INT, both with a positive mode and in increasing order. In particular, a Wilcoxon ranksum test between HI-CSE and HI-INT showed that these distributions had true different locations ($W = 363280$, $p < 0.001$) with HI-CSE segments occurring about 18 ms earlier than HI-INT segments relative to vowel onsets.

4. Discussion

The current study had two objectives: first, to validate the close link between the CSE metric and sound intensity; and second, to examine the relative timing of intensity and CSE-based metrics to explain possible differences. In line with previously reported results (Oxenham *et al.*, 2017; Shu *et al.*, 2016), we found that the CSE metric computed on linearly scaled intensity values of auditory roex filter outputs follows closely the variations of a simple sound intensity characterisation of the speech signal and, as such, speech segments selected by local extrema of the two metrics overlap most of the time. Further, an implementation of the CSE metric employing a more widely used intensity scaling as a precursor to computing spectral change, namely, the auditory-transformed gammatone filterbank and log-transformed decibel intensity values did not predict sentence intelligibility, and in fact had the same impact on intelligibility as a random selection of segments. The effect of the choice of scale, also observed by Oxenham *et al.* (2017), further confirms that spectral change, as measured by the first-order derivative of some energy measure of the speech signal, is not the main determinant of speech segments important for speech recognition. In fact, highly transitional speech portions as selected by the HI-CSE-γ metric (visible in Fig. 1) are not the most intense ones, and therefore are susceptible to noise-masking in real-life environments. Indeed, it is the case that intensity, a first-order indicator of robustness in the face of energetic masking, seems to be a more efficient predictor of intelligibility than a measure of spectral change taken across the full frequency range of auditory speech perception (see Fig. 1).

Despite the close link between CSE and sound intensity, we found a small but significant difference in intelligibility disruption, effectively making the CSE index superior to other signal characterisations of speech segments as measured by the disruptive comprehension paradigm, and as claimed by Stilp and Kluender (2010). This trend is also visible in Fig. 2 of Oxenham *et al.* (2017). We found a systematic shift in the temporal alignment of HI-CSE and HI-INT segments relative to nearby vowel onsets, with CSE peaks occurring slightly earlier than intensity peaks. It appears that the greater disruptive power of replacing HI-CSE segments might be explained by a greater overlap on average with vowel onsets compared to HI-INT segments (see upper panel of Fig. 3). Further studies are needed to test the disruptive power of systematically varying segment selection in relation to the temporal location of intensity peaks. The importance of consonant-vowel transitions, centred on vowel onsets, has also been tested (Fogerty and Kewley-Port, 2009), with the CSE index argued to be a superior predictor (Stilp and Kluender, 2010), although a direct comparison of CSE vs linguistically based segment selection has not been tested. The temporal locus between vowel onset and peak intensity has also been identified as a crucial time instant in rhythmic descriptions of speech perception. The perceptual center, or p-center (Morton *et al.*, 1976) has been defined as the "perceptual moment of occurrence" of a speech unit and

represents the time instant at which listeners would place a beat associated with the speech unit. The temporal midpoint of the rise of amplitude following vowel onset has been proposed as a reasonable acoustic correlate of the p-center, with other factors such as syllable structure and speaking rate also playing a role (Scott, 1993; Cummins and Port, 1998). It is perhaps not surprising that replacing this time instant with noise, as was effectively done in the current study, is highly disruptive, as a consequence of both masking robust bottom-up acoustic cues to speech segmentation and interfering with higher-level rhythmic aspects of speech processing.

5. Conclusion

The current study adds to the evidence of recent investigations using different sets of sentences and languages which together provide a coherent picture of the CSE metric. The metric relies on a non-classical computation of spectral change, based on linear rather than logarithmic (decibel) scaling of intensity values. As such, it is closely related to sound intensity, but with a marked temporal difference: CSE peaks occur significantly earlier than intensity peaks in English. The disparity in temporal alignment may be associated with a small but significant advantage in capturing important speech information, since replacing high-CSE segments with noise disrupts intelligibility to a greater extent than when replacing high-intensity segments.

Acknowledgments

This work was supported by the Australian Reasearch Council under Grant Agreement No. DP150104600 and the European Research Council under the European Community's Seventh Framework Program [FP7/2007-2013 Grant Agreement No. 339152, "Speech Unit(e)s," J.-L. Schwartz PI]. We thank C. Stilp for providing code for the roex filter outputs used in the CSE calculation and H. Faniad for help in collecting data.

References and links

Aubanel, V., Davis, C., and Kim, J. (**2017**). "The MAVA corpus," [online resource] http://dx.doi.org/10.4227/139/59a4c21a896a3 (Last viewed 30 March 2018).

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.

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**(5), 4104–4113.

Cummins, F., and Port, R. F. (**1998**). "Rhythmic constraints on stress timing in English," J. Phon. **26**(2), 145–171.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**(2), 847–857.

Hothorn, T., Bretz, F., and Westfall, P. (**2008**). "Simultaneous inference in general parametric models," Biometr. J. **50**(3), 346–363.

Jiang, Y., Stilp, C. E., and Kluender, K. R. (**2012**). "Cochlea-scaled entropy predicts intelligibility of Mandarin Chinese sentences," in *Proceedings of Meetings on Acoustics*, ASA, p. 060006.

Morton, J., Marcus, S., and Frankish, C. (**1976**). "Perceptual centers (P-centers)," Psychol. Rev. **83**(5), 405–408.

Oxenham, A. J., Boucher, J. E., and Kreft, H. A. (**2017**). "Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy," J. Acoust. Soc. Am. **142**(3), EL264–EL269.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS final report: The Auditory filterbank," Technical Report No. 2341.

Patterson, R. D., Nimmo-Smith, I., Weber, D., 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.

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weistock, M., McGee, V. E., Pachl, U. P., and Voiers, W. D. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Acoust. **17**, 225–246.

Scott, S. K. (**1993**). "P-centers in speech: An acoustic analysis," Ph.D. thesis, UCL, London, UK.

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**(6), EL517–EL521.

Stilp, C., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. USA **107**(27), 12387–12392.