Published in final edited form as:

J Speech Lang Hear Res. 2012 December ; 55(6): 1802–1806. doi:10.1044/1092-4388(2012/11-0255).

The consonant-weighted Envelope Difference Index (cEDI): a proposed technique for quantifying envelope distortion

Eric C. Hoover and Pamela E. Souza

Communication Sciences and Disorders, Roxelyn and Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

Frederick J. Gallun

National Center for Rehabilitative Auditory Research, Department of Veterans Affairs, Portland, OR

Abstract

Purpose—The benefits of amplitude compression in hearing aids may be limited by distortion resulting from rapid gain adjustment. In order to evaluate this, it is convenient to quantify distortion using a metric that is sensitive to the changes in the processed signal that decrease consonant recognition, such as the Envelope Difference Index (EDI). However, the EDI relies on the entire duration of the signal, including portions irrelevant to consonant recognition.

Method—This note describes a computationally efficient method of automatically segmenting speech in time according to the syllable structure. Our technique uses the first derivative of the envelope as a basis. Peaks located in the derivative were used to generate a weighting function for the computation of a metric of signal distortion.

Results—The weighting function significantly improved the variance explained in consonant recognition scores over previous methods, however only 3.2% of the variance was explained in the revised model.

Conclusion—This technique was effective in focusing the analysis of distortion on specific segments of the signal. Use of the technique has implications for speech analysis. The difference in the amplitude envelope of consonants is not a robust model of the effect of hearing aid compression on consonant recognition.

Amplitude compression is widely used in hearing aids to provide audibility without making loud sounds uncomfortable. A negative consequence of amplitude compression is distortion of the temporal envelope (e.g. Jenstad & Souza, 2005), which consists of slow changes in level over time useful for speech understanding. In previous work (Jenstad & Souza, 2005; Jenstad & Souza, 2007; Souza, Hoover, & Gallun, 2012), envelope distortion was measured using the envelope difference index (EDI; Fortune, Woodruff, & Preves, 1994). The EDI is the mean absolute difference in envelope between two signals. For amplitude compression, there is an approximate threshold of EDI=0.2 above which consonant recognition, in both

Copyright 2012 by The American Speech-Language-Hearing Association

Correspondence for this article should be addressed to Eric Hoover, Communication Sciences and Disorders, Northwestern University, Evanston, IL 60201. EricHoover2014@u.northwestern.edu.

Publisher's Disclaimer: This is an author-produced manuscript that has been peer reviewed and accepted for publication in the *Journal of Speech, Language, and Hearing Research (JSLHR)*. As the "Papers in Press" version of the manuscript, it has not yet undergone copyediting, proofreading, or other quality controls associated with final published articles. As the publisher and copyright holder, the American Speech-Language-Hearing Association (ASHA) disclaims any liability resulting from use of inaccurate or misleading data or information contained herein. Further, the authors have disclosed that permission has been obtained for use of any copyrighted material and that, if applicable, conflicts of interest have been noted in the manuscript.

nonsense syllables and sentences decreases (Jenstad & Souza, 2007). This information can be used in the design and prescription of hearing aid compression parameters as an upper limit of acceptable envelope distortion.

A limitation to the conventional EDI when used with vowel-consonant-vowel (VCV) stimuli is that the calculation included portions of the vowels that were unlikely to contribute to consonant recognition. We hypothesize that calculating the EDI from the entire duration of the speech token included envelope distortions irrelevant to consonant recognition. By restricting the analysis to the portion of the token thought to contribute to information transmission, we expect to improve the predictive power of the EDI for levels of signal distortion typical of hearing aid compression.

In order to focus our analysis on the consonant we must make choices about what portions of the signal contribute to consonant recognition. In running speech, this issue is complex (e.g. Drullman, Festen, & Plomp, 1994). In more strictly controlled stimuli such as VCVs, we can reasonably assume that the signal amplitude at the onset of the initial vowel and the offset of the final vowel contribute minimally or not at all to the recognition of the medial consonant. These onsets and offsets can be greatly affected by compression (e.g. overshoot at a vowel onset). We would like to reduce the influence of these regions to our measure of consonant distortion by decreasing the weighting of amplitude changes in regions far from the medial consonant.

Current techniques for automatically parsing syllables attempt to identify vowels at syllable nuclei using peaks in the temporal envelope (Wang & Narayanan, 2007; de Jonge & Wempe, 2009). While this is a robust approach, it is not suitable for our purposes because it is temporally imprecise and susceptible to amplitude compression. Temporal imprecision results from the use of very low envelope filter cutoffs and correlations over time to improve the accuracy of syllable identification in spontaneous speech. Our restricted set of VCV stimuli allow for the use of a less complex method. However, our test stimuli are subjected to amplitude compression, which flattens the temporal envelope and can introduce artificial peaks when gain is reduced (reviewed in Souza, 2002). Because we need to parse speech stimuli that may be heavily compressed, existing techniques relying on locating peaks in the temporal envelope are not appropriate.

Instead of peaks in the temporal envelope, we propose relying on peaks in the first derivative of the temporal envelope. These peaks, both positive and negative, correspond to transitions into and out of pauses in speech, and to the opening and closure of articulators during transitions between consonants and vowels. This is a modification of existing techniques, resulting in a representation of syllables that is temporally precise, informative of both syllable onsets and offsets, and robust to the temporal envelope distortions introduced by amplitude compression.

Algorithm for Syllable Segmentation

The proposed algorithm, the consonant-weighted Envelope Difference Index (cEDI), identifies syllables in a speech signal by selecting consecutive positive and negative peaks from the first derivative of the temporal envelope. Figure 1 (top panel) demonstrates the algorithm output at various stages for an input signal with two syllables. The waveform for the syllable [ama] is shown by the light etching. The envelope (dark solid line) is extracted using rectification and low-pass filtering. The result is down-sampled and the first derivative (medium-dark line) is computed. Peaks in the derivative (vertical dashed lines) are identified that occur at magnitudes greater than a threshold parameter. The highest magnitude peak is selected in alternating directions. These peaks represent the transitions in the envelope from low to high or high to low amplitude, corresponding to the onset and offset of vowels in

each syllable. The bottom panel in Figure 1 is discussed in *Analysis of Consonant Distortion* below.

Methods

The cEDI was evaluated by retrospectively applying it to a subset of previously collected data. Details are available in Souza, Hoover, and Gallun (2012). Briefly, 10 listeners (mean age 25.1 years) with normal hearing (PTA > 25 dB HL) were presented vowel-consonant-vowel (VCV) stimuli, processed to reduce spectral cues so that recognition relied primarily on the low frequency temporal envelope. The stimuli included the consonants /b, d, g, p, t, k, f, θ, s, ʃ, v, ð, z, ʒ, m, n/ spoken with the vowel /a/ by four American English talkers (Turner, Souza, & Forget, 1995). Each stimulus waveform was band limited to 176–7168 Hz, and converted into signal correlated noise (SCN; Schroeder, 1968) by filtering into four bands (crossover frequencies 440, 1130, and 2800 Hz), multiplying each sample of the four time-domain waveforms randomly by either +1 or -1, and filtering the four resulting signals into their respective input bandwidths. Amplitude compression was applied to the stimuli using a hearing aid simulator (GennEM v1.0; Armstrong, 1997) in four conditions determined in a preliminary analysis to produce a range of EDI. The compression ratio and release time for each condition are listed in Table 1, as well as the mean EDI for each condition. Listeners were seated in a sound treated booth and data were collected with a touchscreen monitor. Sounds were presented at 65 dB SPL using Etymotic Research ER-2 insert earphones. Listeners were instructed to select from a grid of consonants the one most similar to the sound presented in a 16-alternative forced-choice paradigm.

Analysis of Consonant Distortion

In order to measure only the envelope changes that could affect consonant recognition, we first identified the segment of each token surrounding the consonant, as follows (see Figure 1, top panel). The VCV was rectified using half-wave rectification and low-pass filtered using a 5th order Butterworth filter with a 50 Hz cutoff frequency. The resulting envelope signal was down-sampled to a sampling rate of 450 Hz. The first derivative of the envelope was computed by the difference between consecutive samples. Peaks greater than a threshold slope parameter of $m=\pm 0.4$ ¹ were labeled. From this set, four peaks were selected as the onset and offsets of the nuclei in each syllable. The peaks were selected using the following logic: for each syllable, the greatest positive peak is the onset until a negative peak is reached, then the most negative peak is the offset until a positive peak is reached indicating the next syllable. The identified locations were inspected manually to ensure the algorithm was successful in all cases.

A window function was computed from the selected peaks, and this window was applied as a scaling function in the EDI computation, resulting in a consonant-weighted metric: cEDI. The window was formed independently for each consonant and talker, based on the syllable landmarks identified from peaks in the derivative of the envelope. The window was a cosine-squared ramp from zero before the first peak to one at the midpoint in time between the first and second peak, then back to zero from the midpoint between the third and fourth peak to the fourth peak. An example window is shown by the thin line in the upper panel of Figure 1.

The computation of the cEDI is as follows:

¹An appropriate threshold can be determined from the peak slope value observed or the expected total number of transitions.

$$cEDI = \frac{1}{2N} \sum_{n=1}^N |w(n) \cdot Env1_n - w(n) \cdot Env2_n| \quad EQ1$$

The absolute value of the difference between the compressed and uncompressed envelopes (Env1 and Env2, normalized) is averaged over each sample (N) and divided by two (Fortune, Woodruff, & Preves, 1994). The computation of the cEDI multiplies each envelope by the window computed above, $w(n)$, prior to normalization. A window of 1 for all n was used when calculating the EDI by the conventional method.

The bottom panel of Figure 1 shows the uncompressed (solid line) and compressed (dotted line) envelopes for /ama/. The grey vertical bars in the lower panel of Figure 1 show the difference values included in the conventional EDI and the darker vertical bars in the lower panel of Figure 1 show the difference values included in the cEDI, resulting from the application of the window function. The window reduced the contribution of the initial vowel onset and final vowel offset, which are regions distorted by compression yet unlikely to contribute to consonant recognition, while not affecting the EDI proximal to the consonant transition regions.

Results

The effect of distortion on temporal envelope cues was quantified using the conventional EDI, as reported previously, and using the cEDI. Figure 2 shows the correlation between the EDI and cEDI for each stimulus waveform in the study. Results from normal hearing listeners recognizing the spectrally-reduced and compressed consonants are shown in Figure 3. Mean consonant recognition was reduced in conditions 3 ($t_g=2.6$, $p=.028$) and 4 ($t_g=4.0$, $p=.003$), where envelope distortion from compression was the greatest. To evaluate our prediction that focusing the analysis of distortion on the consonant portion of the signal would improve its correlation to behavioral performance, both the EDI and cEDI were used in a linear model of consonant recognition. A stepwise linear regression was used with three factors: a metric of distortion (EDI or cEDI), the consonant, and an interaction between consonant and distortion metric. Both the EDI and cEDI accounted for a small but significant portion of the variance in consonant recognition (EDI: $r^2=0.019$, $p=0.028$, cEDI: $r^2=0.032$, $p=0.004$), and the cEDI accounted for more of the variance. The consonant and consonant interaction factors were not significant for either model.

Discussion

Comparison of EDI and cEDI

The results confirm our hypothesis that focusing the analysis of distortion in a signal on the consonant would improve the relationship between distortion and consonant recognition. As demonstrated in Figure 2, the cEDI gives a broader range of values for the same compressed signals compared to the conventional EDI. This indicates that the cEDI is more sensitive to the effects of compression on consonant cues in the temporal envelope than the conventional EDI. As a result, more of the variance in consonant recognition is accounted for with the cEDI. However, the total amount of variance explained remains an unremarkable 3.2%. While it is an improvement over the EDI, the cEDI fails to capture the changes to the temporal envelope that are perceptually salient to the listener. A model based on a frequency analysis of the temporal envelope, like the Spectral Correlation Index (SCI), may be a better approach to quantify the effect of envelope distortion (Gallun & Souza, 2008). Future work will apply the syllable parsing presented here to a model of signal distortion based on the spectrum of the envelope.

Syllable Analysis

The algorithm for automatically parsing syllables presented in this paper was highly successful when applied to VCV stimuli. The technique located the peaks corresponding to every VC and CV transition in the set of 16 consonants for all 4 talkers. A single threshold value was needed for the entire set of signals when uncompressed. The technique was also successful in parsing the compressed signals, but high levels of compression required the selection of a lower threshold value for some consonants. The central two points identified by the algorithm, which indicate the location of the consonant, were not affected by compression. Due to the success of the algorithm under heavy compression, we are confident that it will be useful in a broad range of applications.

Future Directions

The technique for syllable segmentation developed for this paper has applications beyond segmenting the two-syllable speech signals analyzed here. Future work will apply the technique to longer utterances, so that a comparison can be made between short-time, phonemic metrics, like the Spectral Correlation Index (SCI; Gallun & Souza, 2008), and metrics that require a long duration input, like the Speech Intelligibility Index (SII; ANSI 3.5 1997). Such comparisons can be made regardless of the source of envelope distortion (e.g. reverberation, noise reduction). Preliminary data indicate that the algorithm successfully locates syllable boundaries in continuous speech. We plan to use these boundaries as a guide to performing short-time signal analyses on continuous signals for comparison to longer duration metrics. The use of the first derivative of the envelope may also provide an advantage over existing techniques for syllable analysis in speech signal processing or applications such as automatic speech recognition (Bartles & Bilmes, 2007).

Acknowledgments

This work was supported by the National Institutes of Health DC0060014 and by the Veterans Administration Rehabilitation Research and Development Service C4963W.

References

- American National Standards Institute. Methods for calculation of the Speech Intelligibility Index (ANSI S3.5-1997). ANSI; New York: 1997.
- Armstrong, S. GennEm—Single Channel AGC Emulator (Version 1.0) [Computer software]. Gennum; Burlington, Ontario, Canada: 1997.
- Bartels, CD.; Bilmes, JA. Use of syllable nuclei locations to improve ASR. Automatic Speech Recognition & Understanding, 2007, IEEE Workshop on; 2007. p. 335-340.
- de Jong NH, Wempe T. Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods. 2009; 41:385–390. [PubMed: 19363178]
- Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. Journal of the Acoustical Society of America. 1994; 95:2670–2680. [PubMed: 8207140]
- Fortune TW, Woodruff BD, Preves DA. A new technique for quantifying temporal envelope contrasts. Ear and Hearing. 1994; 15:93–99. [PubMed: 8194684]
- Gallun FJ, Souza PE. Exploring the role of the modulation spectrum in phoneme recognition. Ear and Hearing. 2008; 29:800–813. [PubMed: 18596640]
- Jenstad LM, Souza PE. Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility. Journal of Speech, Language, and Hearing Research. 2005; 48:651–667.
- Jenstad LM, Souza PE. Temporal envelope changes of compression and speech rate: combined effects on recognition for older adults. Journal of Speech, Language, and Hearing Research. 2007; 50:1123–1138.

- Schroeder MR. Reference signal for signal quality studies. *Journal of the Acoustical Society of America*. 1968; 104:2467–2476.
- Souza PE. Effects of compression on speech acoustics, intelligibility, and sound quality. *Trends in Amplification*. 2002; 6:131–165.
- Souza PE, Hoover EC, Gallun FJ. Application of the envelope difference index for spectrally sparse speech. *Journal of Speech, Language, and Hearing Research*. 2012 Advance online publication. doi:10.1044/1092-4388(2011/10-0301).
- Wang D, Narayanan S. Robust speech rate estimation for spontaneous speech. *Audio, Speech, and Language Processing, IEEE Transactions on*. 2007; 15:2190–2201.

\$watermark-text

\$watermark-text

\$watermark-text

Figure 1.

[Top panel] Demonstration of the computational steps in determining the transition regions of vowel onsets and offsets. The source waveform (light gray solid line) is rectified and low-pass filtered to produce the envelope (thick, black solid line). The first derivative (dark grey solid line) is computed from the difference between consecutive points in the down-sampled envelope (in this plot normalized to unity). Positive and negative peaks with a magnitude greater than a threshold value are labeled (vertical dotted lines). From these peaks, a window (medium gray solid line) is generated for use in weighting the metric of consonant distortion over time. [Bottom panel] The application of windowing to the Envelope Difference Index (EDI). The envelope of an uncompressed stimulus waveform (thick, black solid line) is compared to the envelope of the same waveform after compression amplification (thick, black dotted line). The absolute value of the difference between these curves is shown as sampled at regular intervals (gray vertical bars). These values are averaged to compute the EDI. The window function (top panel) is applied to the difference values, shown as sampled at regular intervals (black vertical bars). These values are averaged in the cEDI. Note that differences in the envelope at the onset of the initial vowel and the offset of the final vowel are reduced in the cEDI computation.

Figure 2.

Comparison of the EDI to the cEDI. Each circle represents a stimulus waveform used in Souza, Hoover, and Gallun (2012). The effect of compression on the temporal envelope was computed with reference to the uncompressed envelope using the EDI (x-axis) and the cEDI (y-axis). Note that the cEDI occupies a wider range of values.

Figure 3.

Mean and standard error consonant recognition scores for each compression condition. Scores begin to decline at compression condition 4. When listeners are forced to rely on the temporal envelope, distortion resulting from compression can decrease consonant recognition. Data originally published in Souza, Hoover, and Gallun (2012).

Table 1

Compression parameters and mean EDI for each compression condition. An attack time of 6 ms and a compression threshold of 40 dB SPL was used in each condition.

Condition	Compression Ratio	Release Time (ms)	Mean EDI
0	1:1	n/a	0
1	2:1	800	0.06
2	4:1	800	0.11
3	4:1	12	0.18
4	10:1	12	0.23