Published in final edited form as:

Proc Meet Acoust. 2013 June ; 19: 060118–060124.

When Spectral Smearing Can Increase Speech Intelligibility

J.A. Bashford, R.M. Warren, and P.W. Lenz

POMA of poster presented at the ASA/ICA 2013 meeting in Montreal, CA

Abstract

Sentences were reduced to an array of sixteen effectively rectangular bands (RBs) having center frequencies ranging from 0.25 to 8 kHz spaced at $\frac{1}{3}$ -octave intervals. Four arrays were employed, each having uniform subcritical bandwidths which ranged from 40 Hz to 5 Hz. The 40 Hz width array had intelligibility near ceiling, and the 5 Hz array about 1%. The finding of interest was that when the subcritical speech RBs were used to modulate RBs of noise having the same center frequency as the speech but having bandwidths increased to a critical (ERBn) bandwidth at each center frequency, these spectrally smeared arrays were considerably more intelligible in all but the 40 Hz (ceiling) condition. For example, when the 10 Hz bandwidth speech array having an intelligibility of 8% modulated the ERBn noise array, intelligibility increased to 48%. This six-fold increase occurred despite elimination of spectral fine structure and addition of stochastic fluctuation to speech envelope cues. (As anticipated, conventional vocoding with matching bandwidths of speech and noise reduced the 10-Hz-speech array intelligibility from 8% to 1%). These effects of smearing confirm findings by Bashford, Warren, and Lenz (2010) that optimal temporal processing requires stimulation of a critical bandwidth. [Supported by NIH]

INTRODUCTION

Previous studies have shown that speech is highly redundant in the spectral domain, remaining intelligible despite substantial reductions in bandwidth (e.g., Warren et al., 1995) and/or the removal of spectral fine structure (e.g., Shannon, et al. 1995). Bashford, Warren, and Lenz (2010) examined the lower limits of spectral redundancy using high-order FIR filtering to reduce broadband “everyday” sentences to an array of 16 rectangular bands, each having subcritical bandwidths. Generation of these novel speech bands requires extremely steep, effectively vertical filter slopes: Warren, Bashford, and Lenz (2004) found that transition bands having a 1200 dB/octave roll-off contributed over 30% of the intelligibility of a nominal $\frac{1}{3}$ -octave passband, and they found that intelligibility of the speech band did not reach its minimum until roll-off was increased to 4800 dB/octave. Bashford et al. (2010) maintained that required roll-off across their 16-band speech arrays through frequency-dependent adjustments in the order of FIR filtering, and they examined the intelligibility of arrays comprised of bands having widths of 4%, 2%, 1% and 0.5% relative to their center frequencies. Nearly perfect sentence intelligibility (96%) was obtained for the array comprised of 4% bands, which was an impressive demonstration of spectral redundancy, given that filtering removed approximately 90% of the speech signal across the audiometric range from 125 Hz to 8000 Hz. An array comprised of 2% bands, having about 95% of the speech signal removed, remained 80% intelligible. In the 1% bandwidth condition intelligibility remained appreciable (25%), and although near floor (2.5%) in the 0.5% bandwidth condition (2.5%), it did differ significantly from zero [$t(19) = 3.66$, $p < 0.003$]. These results indicate that natural speech, containing both spectral fine structure and temporal envelope information, can remain highly intelligible when reduced to an exceedingly sparse spectral representation. However, as discussed below, when Bashford et al. (2010) removed fine-structure information by noise-vocoding their speech band arrays, they found that effective use of the remaining, temporal envelope cues was seriously compromised when bandwidths were extremely narrow.

In order to examine the effects of subcritical-band filtering on the processing of temporal cues, Bashford et al. (2010) extracted the envelopes from their speech band arrays through full-wave rectification and used those envelopes to modulate 16 rectangular bands of white noise that matched the speech in bandwidth, ranging from 4% to 0.5%. As expected, intelligibility of the noise-vocoded arrays relative to the original speech bands was much lower when the noise bands matched the subcritical widths of the parent speech bands. This was most dramatic for 2% bands, with intelligibility declining from a value of 80% for the speech bands to only about 20% for the array comprised of corresponding speech-modulated noise bands. This very large loss of intelligibility produced by the removal of fine-structure cues, suggests that the normally dominant contribution of envelope cues to intelligibility (e.g., Smith, Delgutte, & Oxenham, 2002) may require the stimulation of a majority of the fibers comprising individual critical bands. Support for this hypothesis was provided by two additional noise vocoding conditions: Somewhat surprisingly, intelligibility rose to about 85% when the envelope information extracted from the 2% speech bands was used to modulate arrays of either ERBn- or 1/3-octave-wide noise bands.

The present study was designed to extend these findings to conditions permitting an examination of the effects of both bandwidth and envelope frequency range upon temporal processing. Since the maximum rate of envelope fluctuation for a speech band (or a modulated noise band derived from it) is equal to the band's width in Hz, the upper limit of envelope fluctuation increased with speech band center frequency in the Bashford et al. (2010) study: For example, in the 4% bandwidth condition the limiting envelope frequency ranged from 10 Hz for the speech band centered at 250 Hz to 320 Hz for the highest band centered at 8000 Hz. The present study examined the intelligibility of speech and modulated noise band arrays based upon initial subcritical bandpass filtering of broadband sentences in fixed steps of 40, 20, 10 or 5 Hz. As was the case for the fixed percentage steps used earlier, this range in Hz-wide steps allowed performance to range from near ceiling to near floor intelligibility.

METHODS

Listeners

The 80 listeners in this study (4 groups of 20) were undergraduate students at the University of Wisconsin-Milwaukee who were paid for their participation. All listeners were native monolingual English speakers who reported having no hearing problems and had normal bilateral hearing, as measured by pure tone thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz.

Stimuli

The speech stimuli were the 100 (10 lists of 10) CID "everyday" sentences that contain 500 keywords used for scoring (Silverman & Hirsh, 1955). They were derived from a broadband digital recording (44.1 kHz sampling, 16-bit quantization) used previously by Warren et al. (1995) in a preliminary study of spectral redundancy. The recording was produced by a male speaker having no evident regional accent and an average voicing frequency of 100 Hz.

Figure 1 shows the long-term averaged spectra of the original broadband sentences (in green), along with the spectrum of sixteen rectangular bands (shown in yellow) extracted from the broadband stimulus through FIR filtering. The bands shown have a fixed bandwidth of 2%, and center frequencies ranging from 250 Hz to 8000 Hz, spaced at 1/3-octave intervals. These subcritical width bands were then combined with their spectrum levels equated for use in listening experiments.

In FIR filtering, a given number of taps will produce a roll-off that is fixed in dB/Hz. Hence, in order to produce a roll-off fixed in dB/octave it is necessary to employ a proportionally greater number of taps as the cutoff frequency decreases. In order to produce the 4800 dB/octave roll-off required for this study, the number of filter taps ranged from a low of 1404, required for the higher frequency skirt of the 8000 Hz CF band, to a high of 382,000 taps required for the lower frequency skirt of the 250 Hz CF band. Filtering of the sentences was accomplished using the *fir 1* function in MATLAB 7.7.0.471 running on a Mac Pro 2 X 2.93 GHz quad-core Intel Xeon processor.

Because effective subcritical width filtering employs exceedingly high orders of filtering to produce sub-critical bandwidths, antiphase cancellation was used to examine the stimuli for possible artifacts or signal distortions. Using the same high order of filtering employed in the present study, Bashford et al. (2010) extracted from the broadband signal sixteen contiguous 1/3-octave speech bands which were added together, phase inverted, and then added back to the broadband signal to determine what artifacts remained after cancellation. Figure 2 shows the spectrum of the original signal (in green) along with a tracing of spectral components that remained in the range of the added antiphase passbands after cancellation (yellow tracing). It can be seen that cancellation within the passbands was approximately 60 dB, indicating negligible distortion of the speech signal. Moreover, the ringing components at the cutoff frequencies (yellow spikes), produced by FIR filtering, were more than 30 dB down for each of the speech bands.

In the present study, one group of listeners was presented with 16-band speech stimuli varying in bandwidth from 40 Hz to 5 Hz. Using the same procedure, the remaining three groups of listeners were presented with 16-band arrays consisting of rectangular noise bands that were modulated by the temporal envelopes of corresponding speech bands extracted by full-wave rectification. The vocoded noise bands were filtered either to match the bandwidth of their parent speech bands, or to exceed that bandwidth, with expansion to either 1 ERBn or 1/3-octave.

PROCEDURE

The four groups of listeners who participated in this study received 5 blocks of twenty sentences. The first two blocks of sentences presented were derived from 40-Hz speech bands, with the first block serving as practice. The third, fourth and fifth blocks of sentences were derived from 20, 10, and 5 Hz bands, respectively. As in the earlier study (Bashford et al., 2010), stimuli in this experiment were presented in order of decreasing speech bandwidth to allow listeners to adapt to the effects of filtering. One group of listeners received the 16-band speech stimuli, while the remaining three groups received 16-band vocoded stimuli whose component noise bands were modulated by the temporal envelopes extracted from the speech bands having corresponding center frequencies. For one of the latter groups, the vocoded noise bands matched the bandwidths of the parent speech bands. For the remaining two groups the noise bands were expanded beyond the spectral limits of the speech bands, either to the width of an ERBn or to 1/3-octave.

Testing was performed in a sound-attenuating chamber, with the stimuli delivered diotically through Sennheiser HD 250 Linear II Headphones at a slow-rms peak level of 60 dBA SPL. Listeners were instructed to call out what the voice was saying as best they could, and were encouraged to guess when unsure.

RESULTS

Figure 3 presents the mean percent intelligibility scores obtained for the four groups of listeners, plotted as a function of speech bandwidth. For comparison, Figure 4 presents the

corresponding results obtained by Bashford et al. (2010) for bands varied in percentage steps. It can be seen in Figure 3 that performance for listeners receiving the filtered speech bands was near ceiling (95%) for the second block of sentences comprised of 40-Hz bands. Intelligibility declined to about 50% for the 20-Hz bands, and fell to near zero when the widths of the speech bands were decreased to 5 Hz. For the remaining groups of listeners, who received noise-vocoded stimuli, performance varied dramatically as a function of noise bandwidth. For listeners receiving the noise bands that matched the subcritical bandwidths of the parent speech bands, intelligibility dropped substantially relative to that obtained with the speech, and most dramatically in the 40-Hz bandwidth condition, with an intelligibility decrease of 80% and 70% in the first and second blocks of that condition, respectively. However, this loss of intelligibility did not occur when the envelopes of the 40-Hz-wide speech bands modulated either ERBn or 1/3-octave noise bands. Intelligibility for those arrays was statistically equivalent to that of the parent speech array. Consistent with the findings of Bashford et al. (2010), in the remaining bandwidth conditions intelligibility was higher for the ERBn and 1/3-octave vocoded stimuli than for the speech, indicating that optimal envelope processing requires stimulation of the majority of fibers comprising critical bands. When that requirement was met, as shown in Figure 3, starting with the second block of sentences variation of envelope frequency produced a linear decrease of intelligibility across the three octaves from 40 Hz to 5 Hz.

DISCUSSION

Intelligibility scores were somewhat lower in most conditions of this experiment than those obtained by Bashford et al. (2010), and they were substantially lower for the speech-modulated noise bands that were presented at the subcritical widths matching those of the filtered speech bands. This difference was anticipated based upon the differential effects of fixed-percentage- vs. fixed-Hz-width bandpass filtering in the mid and high frequency regions of the speech signal. With bandpass filtering at a 40-Hz bandwidth, for example, the 9 bands above 1000 Hz, ranging from 1250 Hz to 8000 Hz, ranged from 25% to 4.5% of an ERBn in width, whereas bands having the same center frequencies produced by 4% bandpass filtering ranged from 31% to 36% of an ERBn. Hence, the greater intelligibility of the 4% speech and modulated noise bands is consistent with the hypothesis that the effectiveness of temporal processing increases with the proportion of nerve fibers recruited within individual critical bands. These findings also converge with additional evidence obtained by Bashford et al. (2010), (Experiments 2 and 3 of that study) who extracted the envelopes of 2% speech bands and used them to modulate an array of noise bands that varied from 5% to 300% of an ERBn. They obtained a linear intelligibility increase over the range from 5% to 80% of an ERBn, reaching an initial asymptotic intelligibility of approximately 85% that extended to 150% ERBn width. There was a gradual decline of intelligibility that was significant for the 200% ERBn condition and wider bandwidths ($p < 0.05$ by Tukey test).

Considered together the results of these two studies employing effectively subcritical band speech clearly show that envelope processing is obligatorily coarse-grained, and this limitation may help explain some earlier observations in the literature involving spectrally sparse analogs of natural speech. For example, Souza and Rosen (2009) have reported higher intelligibility for noise-vocoded than sine-vocoded speech when extracted speech envelopes were subjected to 30 Hz rather than 300 Hz lowpass filtering. They suggested that the greater “density” of the noise spectrum under these filtering conditions might facilitate auditory object formation. Alternatively, our results suggest that the greater density of the noise spectrum improves the processing of envelope cues within critical bands, prior to object formation. The present findings also can explain how the intelligibility of sine-wave speech can be improved by amplitude modulation of the sine-waves that track speech

formants (Carrell & Opie, 1992; Lewis & Carrell, 2007). Results obtained with subcritical speech band arrays suggest that the broadening of sine-wave formant trackers through amplitude modulation may improve intelligibility by more fully engaging coarse-grained processing of temporal envelope cues, which can then operate along with the processing of spectral fine structure.

Acknowledgments

Research reported in this publication was supported by the National Institute On Deafness And Other Communication Disorders of the National Institutes of Health under Award Number R01DC000208. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Bashford JA Jr, Warren RM, Lenz PW. Enhancing the intelligibility of high-speed: Evidence of inhibition in the lower auditory pathway. *Proc Mtgs Acoust.* 2011; 12:1–6.
- Carrell TD, Opie JM. The effect of amplitude comodulation on auditory object formation in sentence perception. *Percept Psychophys.* 1992; 52:437–445. [PubMed: 1437476]
- Lewis DE, Carrell TD. The effect of amplitude modulation on intelligibility of time-varying sinusoidal speech in children and adults. *Percept Psychophys.* 2007; 69:1140–1151. [PubMed: 18038952]
- Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science.* 1995; 270:303–304. [PubMed: 7569981]
- Silverman SR, Hirsh IJ. Problems related to the use of speech in clinical audiometry. *Ann Otol Rhinol Laryngol.* 1955; 64:1234–1245.
- Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. *Nature.* 2002; 416:87–90. [PubMed: 11882898]
- Souza P, Rosen S. Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *J Acoust Soc Am.* 2009; 126:792–805. [PubMed: 19640044]
- Warren RM, Riener KR, Bashford JA Jr, Brubaker BS. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Percept Psychophys.* 1995; 57:175–182. [PubMed: 7885815]
- Warren RM, Bashford JA Jr, Lenz PW. Intelligibility of bandpass filtered speech: Steepness of slopes required to eliminate transition band contributions. *J Acoust Soc Am.* 2004; 115:1292–1295. [PubMed: 15058351]

FIGURE 1.

Long-term averaged spectra based upon the 100 CID “everyday” sentences. The spectrum of the unfiltered broadband sentences is shown in green, along with that of sixteen 2% rectangular bands produced via high-order FIR filtering to produce effectively vertical slopes of 4800 dB/octave or better. Plots were generated via discrete Fourier transform, with linear averaging of spectra obtained using a 8192-point (186 ms, 5 Hz resolution) Blackman window (from Bashford et al., 2010).

FIGURE 2.

Evaluating the fidelity of high-order FIR filtering. A series of sixteen contiguous 1/3-octave bands of speech having the same center frequencies as the experimental stimuli were generated via high-order FIR filtering and then were phase inverted and added back to the broadband stimulus, shown in green above. The spectrum of components remaining following antiphasic addition, shown in yellow, reveals that substantial cancellation occurred within the passbands (generally better than 60 dB) and that ringing components at the cutoff frequencies of the bands, produced by FIR filtering, was more than 30-dB down across the spectrum (from Bashford et al., 2010).

FIGURE 3.

Mean percent intelligibility scores for the four types of 16-band stimuli presented to listeners in the present study. One group of listeners received stimuli comprised of rectangular speech bands ranging from 40 Hz to 5 Hz in bandwidth (“Speech” group). For the remaining 3 groups of listeners, who received speech-modulated noise (SMN), the temporal envelopes were extracted from the speech bands and used to modulate 16 rectangular bands of noise that either matched the speech in bandwidth (“Matched BW SMN” group), or were increased either to ERBn width (“ERBn SMN” group), or to a width of 1/3-octave (“1/3-octave SMN” group). Note that there are two entries for the 40 Hz bandwidth sentences: the first entry represents a block of practice trials with a separate block of sentences presented to familiarize listeners with the novel band limited stimuli.

FIGURE 4.

Mean percent intelligibility scores for the 16-band stimuli presented to listeners in Experiment 1 of Bashford et al. (2010). One group of listeners received stimuli comprised of rectangular speech bands ranging from 4% to 0.5% in bandwidth ("Speech" group). For the remaining 3 groups of listeners, who received speech-modulated noise (SMN), the temporal envelopes were extracted from the speech bands and used to modulate 16 rectangular bands of noise that either matched the speech in bandwidth ("Matched BW SMN" group), or were increased either to ERBn width ("ERBn SMN" group), or to a width of 1/3-octave ("1/3-octave SMN" group).