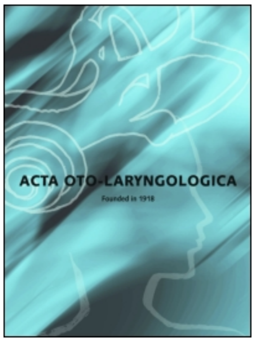

Acta Oto-Laryngologica

ISSN: 0001-6489 (Print) 1651-2251 (Online) Journal homepage: https://www.tandfonline.com/loi/ioto20

# Spectral Enhancement to Improve the Intelligibility of Speech in Noise for Hearing-impaired Listeners

A. M. SimpsonA. M. SIMPSON, B. C. J. MooreB. C. J. MOORE & B. R. GlasbergB. R. GLASBERG

**To cite this article:** A. M. SimpsonA. M. SIMPSON, B. C. J. MooreB. C. J. MOORE & B. R. GlasbergB. R. GLASBERG (1990) Spectral Enhancement to Improve the Intelligibility of Speech in Noise for Hearing-impaired Listeners, Acta Oto-Laryngologica, 109:sup469, 101-107, DOI: 10.1080/00016489.1990.12088415

**To link to this article:** https://doi.org/10.1080/00016489.1990.12088415

Published online: 20 Dec 2019.

# Spectral Enhancement to Improve the Intelligibility of Speech in Noise for Hearing-impaired Listeners

A. M. SIMPSON, B. C. J. MOORE and B. R. GLASBERG

*From the Department of Experimental Psychology, University of Cambridge, England*

Simpson AM, Moore BCJ, Glasberg BR. Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. Acta Otolaryngol (Stockh) 1990; Suppl. 469: 101–107.

At speech-to-noise ratios between −3 and 6 dB, many hearing-impaired listeners have difficulty in understanding speech, but spectrograms reveal that the formant peaks of voiced speech and some of the spectral peaks associated with unvoiced speech stand out against the background noise. Our speech-enhancement process is based on the assumption that increasing spectral contrast will result in improved intelligibility. The enhancement involves calculating an auditory excitation pattern from the magnitude spectrum of overlapping short segments of the speech signal. This pattern is convolved with a difference-of-Gaussians function whose bandwidth varies with frequency in the same way as the auditory filter bandwidth. Magnitude values from this enhanced pattern are combined with the unchanged phase spectrum from the original signal to produce the enhanced speech. The processing was used to enhance Boothroyd and Bench-Kowal-Bamford Audiometric lists which had been digitally combined with speech-shaped noise at speech-to-noise ratios between −3 and 6 dB. The subjects had moderate to severe sensorineural hearing losses. The processing produced small but significant improvements in intelligibility for the hearing-impaired listeners tested. Possibilities for improving the processing are discussed.
*Key words: digital processing of speech.*

## INTRODUCTION

At speech-to-noise ratios (SNRs) between −3 and +6 dB many listeners with a sensorineural hearing loss have difficulty in understanding speech. Even when the overall level of the speech-plus-noise is high enough to overcome any attenuative component of their impairment, noise can obscure the speech to a greater degree than for normal listeners (1, 2). This paper describes a method for the digital processing of speech in noise which holds some promise for alleviating this problem.

Spectrograms reveal that for SNRs between −3 and +6 dB, most of the formants of voiced speech and some of the high-frequency peaks associated with unvoiced speech are visible in noise that has the same long-term-average spectrum as the speech. Normal listeners appear to be able to extract information about the speech from the local frequency regions where the SNR is high, whereas hearing-impaired listeners have difficulty. Somehow the noise between the formant peaks has adverse effects for impaired listeners, possibly because their frequency selectivity is reduced compared with normal (2, 3). If the noise in the spectral valleys could be attenuated, this might improve the intelligibility of the speech for hearing-impaired listeners. In effect, this is what our processing scheme does; it increases the contrast in the short-term spectrum of the speech-plus-noise by enhancing the peaks relative to the valleys.

A number of previous studies have looked at the effect of spectral 'sharpening' on speech intelligibility, mostly with negative results. Boers (4) processed a set of Dutch sentences to increase their spectral contrast. The speech was filtered through twelve one-third octave digital filters. The amplitude of each filter output was multiplied by its absolute value and all of the outputs rescaled, increasing the contrast between peaks and valleys in the spectrum. High-frequency emphasis and amplitude compression were also carried out. Noise was added to the speech *after* processing and speech reception thresholds were compared for processed and unprocessed speech. Processing reduced intelligibility for 4 out of the 6 hearing-impaired listeners tested and all 7 of the normal hearing listeners. The processing gave a slight improvement for 2 of the hearing-impaired listeners.

Summerfield et al. (5) tested the effects of altering formant bandwidths on the identification of synthesized consonant–vowel–consonant (CVC) syllables. Stimuli were presented in quiet. For both normal and impaired listeners, broadening the formants had deleterious effects on identification accuracy. However, narrowing the formants did not improve recognition performance for either group.

There are several possible reasons why these studies failed to show benefits for spectral sharpening. Boers' study did not include any control conditions to show the effect of each stage of the processing. Also, the noise was added after processing, so noise falling in the valleys between formants would not have been reduced in level. Summerfield et al. did not present the stimuli in noise and they only used 'whispered' (noise excited) speech.

Our processing operates to enhance the spectral contrast of speech which has already been contaminated by background noise. One drawback of increasing the spectral contrast of a speech signal in noise is that spectral peaks in the noise may also be enhanced, leading to spurious spectral features which can have a deleterious effect on speech perception. One way of reducing this problem is to 'smooth' the spectrum prior to enhancement. Our method of smoothing is based on the following reasoning. The normal ear is well suited for extracting speech from noisy backgrounds. Thus the processing should avoid enhancing spectral features which would not be resolved by a normal listener; the smoothing of the spectrum should remove those spectral features which would not be resolved by a normal ear, but preserve those features which would be resolved. This can be done by converting the short-term spectrum to an auditory excitation pattern, as described by Moore & Glasberg (6). The spectral enhancement can then be performed on the excitation pattern.

## METHOD OF SPEECH ENHANCEMENT

Our spectral enhancement technique involves manipulating the short-term spectrum of the speech signal. Sampled segments of the speech are windowed, smoothed, spectrally enhanced and then resynthesized using the overlap-add technique (7). The precise steps are as follows:

1) The speech is low-pass filtered at 4 kHz and sampled at a 10-kHz rate with 12-bit resolution.

2) A 25.6-ms segment of speech is weighted by a 25.6-ms Hamming window.

3) A 256-point fast Fourier transform (FFT) of the windowed sample values is calculated, giving 128 magnitude values and 128 phase values. The phase values are stored, and subsequent operations are only carried out on the magnitude spectrum. An example of a magnitude spectrum calculated in this way is shown in the top panel of Fig. 1.

4) To remove any perceptually irrelevant spectral detail, the excitation pattern corresponding to the short-term magnitude spectrum is calculated, using the method described by Moore & Glasberg (6). This involves calculating the output of an array of auditory filters whose bandwidth increases with increasing centre frequency; the excitation pattern is defined as the output of these filters as a function of centre frequency. This produces 128 new magnitude values. An example of such an excitation pattern is shown in the middle panel of Fig. 1. Note that the minor irregularities in the magnitude spectrum (top) have

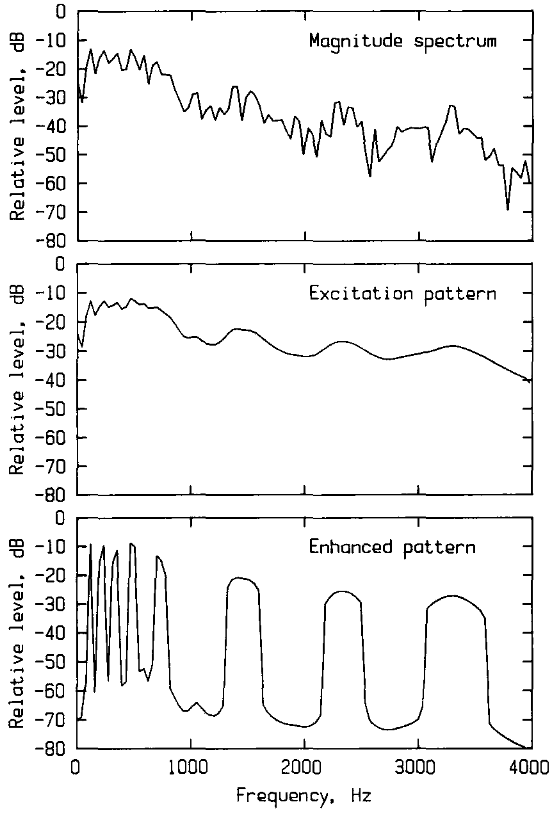

*Fig. 1. (Top panel)* Magnitude spectrum of a segment of speech in noise, obtained via an FFT. *(Middle panel)* Auditory excitation pattern corresponding to that magnitude spectrum. *(Bottom panel)* Excitation pattern after enhancement with the DOG function (see main text for details).

been smoothed out, but the peaks corresponding to the formant frequencies have been preserved.

5) The excitation pattern is enhanced by a process resembling convolution with a difference-of-Gaussians (DOG) function or 'Mexican Hat' function. This function is the sum of a positive Gaussian and a negative Gaussian which has twice the bandwidth of the positive Gaussian, as described by the following equation:

$$\mathrm{DOG}\,(\Delta f) = (1/2\pi)^{1/2}[\exp\{-(\Delta f/b)^2/2\} - (1/2)\exp\{-(\Delta f/2b)^2/2\}],$$

where $f$ is deviation from the centre frequency, and $b$ is a parameter determining the bandwidth of the DOG function. Note that the total area of this function, summed over positive and negative portions, is zero. The value of $b$ for a given centre frequency is chosen so that the width of the positive lobe at the zero-crossing point is made equal to the equivalent rectangular bandwidth (ERB) (6) of the auditory filter with the same centre frequency. Thus the width of the DOG function increases with increasing centre frequency according to a scale representing the frequency selectivity of the ear.

The DOG function is centred on the frequency of each of the 128 magnitude values in turn. For a given centre frequency of the DOG function, the value of the excitation pattern at each frequency is multiplied by the magnitude of the DOG function at that same frequency, and the products obtained in this way are summed. The magnitude value of the excitation pattern at that centre frequency is then replaced by that sum. For frequency regions where the magnitude spectrum is relatively flat, the sum is close to zero. However, where the spectrum contains peaks the sum is positive, and where it contains valleys the sum is negative. The sum is particularly large for peaks surrounded by valleys, especially when the peaks and valleys are of comparable frequency extent to the positive and negative portions of the DOG function.

In order to control the degree of enhancement applied, the output of each DOG filter is scaled and then added to the corresponding magnitude value from the original excitation pattern. An example of the resultant enhanced pattern is shown in the lower panel of Fig. 1. Note that the peaks and valleys appear broad at high frequencies and narrow at low frequencies. This is a consequence of the linear frequency scale used in the diagram. If the patterns were plotted in ERB units (6) the peaks and valleys would appear to be more or less uniform over the whole frequency range.

6) The magnitude values from the enhanced excitation pattern are combined with the unchanged phase spectrum and an inverse FFT is used to produce a 25.6-ms segment of spectrally enhanced speech.

7) The process is repeated every 12.8 ms, and the resultant overlapping enhanced speech segments are summed to give a complete processed speech waveform.

In his original article on the overlap-add method, Allen (7) recommended that windowed segments should be taken at intervals corresponding to one-quarter of the window length. This means that each window overlaps its immediate neighbours by three-quarters of its length, rather than one-half of the window length we used. However, in preliminary experiments we found that overlapping windows by a half as opposed to three-quarters had no perceptible effect for speech in noise. Allen also recommended that the speech segment length should be shorter than the window length, and that the sample values should be padded with zeros prior to performing the FFT. Again, preliminary experiments indicated that padding with zeros had no perceptible effect for speech in noise.

## PREPARING THE STIMULI

Test stimuli were Boothroyd word lists (9), scored by phoneme, and Bench-Kowal-Bamford (BKB) sentence lists (9), scored by key word, presented in continuous background noise. For most of the tests, the background noise for each set of stimuli was digitally synthesized so as to have the same long-term average spectrum as the speech. To determine this spectrum, an analysis of the speech, identical to stages 1–3 of the enhancement process was carried out (i.e. low-pass filtering the speech, producing a Hamming-weighted segment of speech 25.6 ms in length, and using a 256-point FFT to calculate 128 magnitude and phase values). Each magnitude value was squared to give a power spectrum and successive power spectra averaged over the length of several sentence lists to provide a long-term average power spectrum of the speech. This average power spectrum was then converted to an amplitude spectrum by taking the square-root of each magnitude value. The 128 magnitude values were combined with 128 random phase values and a 256-point inverse FFT used to give a 25.6-ms sample of noise. This was Hamming-weighted and overlapped every 12.8 ms with a new 25.6-ms sample of noise calculated from the same magnitude spectrum but with a new random phase spectrum. By repeating this process, noise of sufficient duration for the speech stimuli was produced. Each sentence or word list was then combined with the corresponding noise at SNRs of −3 to 6 dB.

Some tests were carried out using noise whose long-term-average spectrum was similar to that of the speech for frequencies up to 1.5 kHz, but which had progressively less energy at higher frequencies (about 5 dB less at 4 kHz). This noise simulates a typical listening situation where several people are talking at once: the surfaces of rooms tend to absorb the high frequencies more than the low, so that the background noise has relatively less high-frequency energy than the speech.

All lists were processed in two ways. In the first, they were spectrally enhanced as described earlier. In the second (control), the sentences went through all the stages of processing except those designed to smooth and enhance the spectrum (stages 4 and 5). The enhancement process introduced a slight high-frequency emphasis to the signals. However, the long-term average spectrum of the segments containing speech-plus-noise remained the same as that of segments containing noise alone. The high-frequency emphasis was removed by analogue and/or digital filtering. After this filtering the long-term average spectrum of the enhanced and non-enhanced speech in noise was identical within ±1 dB over the range 0–5 kHz.

## TESTING

Nine listeners with moderate to severe hearing losses were used, although some were only tested in a few conditions. All listeners were diagnosed as having a sensorineural hearing loss, and all had loudness recruitment as revealed by a smaller than normal dynamic range between threshold and discomfort level. Three of the subjects (DJ, GB and PM) had a unilateral hearing loss and were tested using their impaired ear only. This was achieved by fitting their normal ear with a silicon ear mould and then presenting pink noise to that ear using a single earpiece of a Beyer DT48 headset. None of the subjects with a unilateral loss wore an aid. Subjects with bilateral losses and who normally wore an aid (AS, EH, IH and RL) were asked to use them during the tests and to adjust them to their optimum listening level. All wore a conventional behind-the-ear aid in one ear only.

Subjects sat in a sound-attenuating room facing a single loudspeaker (monitor audio MA4) at a distance of 1.3 metres. Initially, sentences were presented so that the noise between the speech was at an average level of 75 dB SPL for both enhanced and control stimuli. Subjects were told to try and recognize as many of the words as they could, and to make a guess when they were not sure. They were also told that the task would be quite difficult and that they were not expected to be able to hear every word. Two BKB lists (one enhanced, one control) or one Boothroyd list (enhanced only) were used to familiarize subjects with the task and to allow them to adjust their aid to the optimum level. For the subjects who listened unaided, the noise level was sometimes increased to ensure that the speech and noise were well above their absolute threshold. These initial tests were also used to select the most appropriate speech-to-noise ratio for each subject. After hearing each list subjects were given a one minute rest. Subjects were tested on 12 Boothroyd word lists in one sitting, but were tested on 12 BKB sentence lists in two groups of six lists with a half-hour break in between. Word and sentence list order, and whether a given list was enhanced or not, was randomized so that for a given test session half the lists were enhanced and half were control lists. Subjects were given no feedback as to their performance until the end of the test.

## RESULTS

Fig. 2A shows results using the BKB sentence lists presented in noise with the same long-term average spectrum as the lists. The speech-to-noise ratio was chosen for each subject to try to achieve scores in the range 40–80%; the actual ratio used is shown above each histogram bar. Each of the subjects showed improved intelligibility for the enhanced speech, the improvements ranging from 1 to 10%. The mean improvement was 6.8%. A related-samples $t$-test showed that the improvement was significant at $p<0.005$ (one-tailed test).

Fig. 2B shows results using the Boothroyd word lists presented in noise with the same

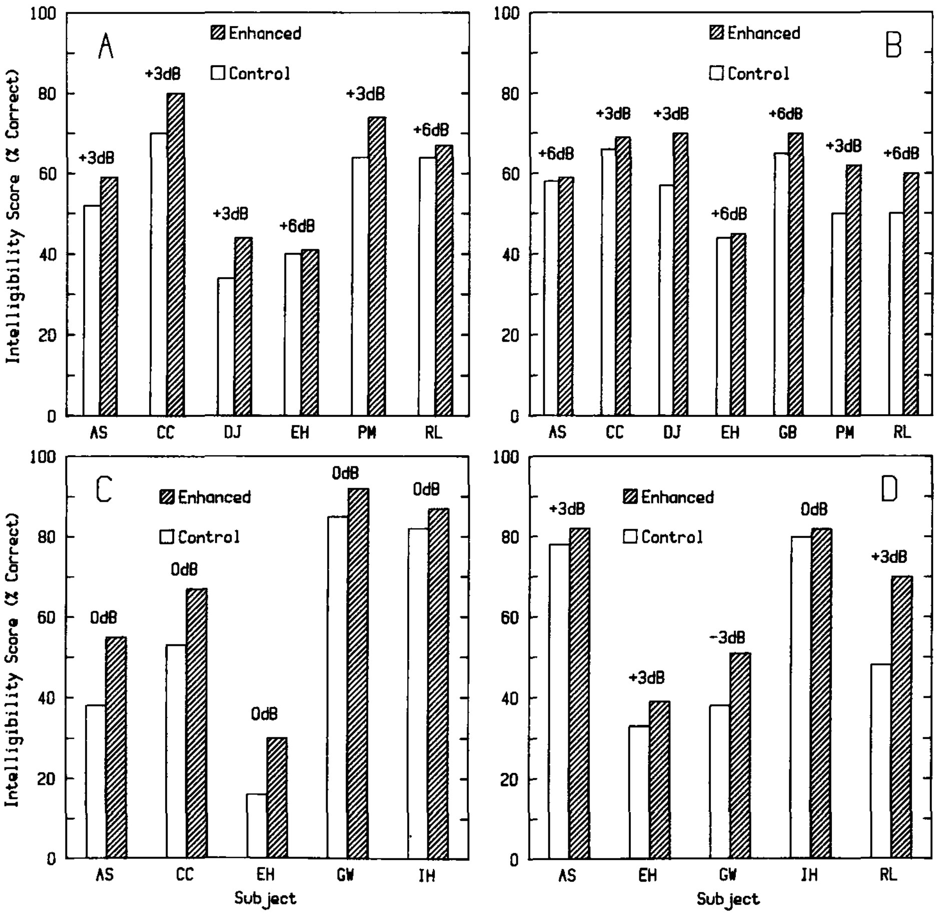

*Fig. 2.* Results of the speech intelligibility tests. Each histogram bar shows results for one subject with control *(open)* or enhanced *(shaded)* speech. Numbers above the bars show the speech-to-noise ratio used. Panels *A, C* and *D* show results using the BKB sentence lists. Panel *B* shows results using the Boothroyd word lists. For panels *A* and *B*, the noise had the same long-term average spectrum as the speech. For panels *C* and *D*, the noise spectrum had relatively less energy at high frequencies (see main text for details).

long-term average spectrum as the lists. Again, the speech-to-noise ratio was chosen separately for each subject. Each of the subjects showed an improvement for the enhanced speech; the mean improvement was 6.4%. A related-samples *t*-test showed that the improvement was significant at $p<0.01$ (one-tailed test). For initial consonants, the processing was most effective for glides and voiced plosives. It was also effective for /f/ and the affricate /dʒ/. For final consonants, the processing was most effective for the unvoiced plosives /t/ and /k/. However, the enhancement process sometimes emphasised acoustic features in a way which made phonetic identity ambiguous. For example, with the word /rap/ the aspiration of the final /p/ was emphasized, making the /p/ sound more like a /t/. The vowels were generally well identified in both the enhanced and non-enhanced speech.

Fig. 2 C shows results for the BKB lists using the noise whose spectrum had slightly less energy at high frequencies than the speech. The speech-to-noise ratio was 0 dB for all subjects. All subjects showed an improvement for the enhanced speech, the mean improvement being 11.4%. A related-samples *t*-test showed that the improvement was significant at $p<0.005$ (one-tailed test). However, some subjects had very low or very high scores overall. To reduce floor and ceiling effects, a second test using this noise was carried out choosing the speech-to-noise ratio separately for each subject to try to achieve scores between 40 and 80% correct. The results are shown in Fig. 2D. Again, each subject showed an improvement, the mean improvement being 9.4%. A related-samples *t*-test showed that the improvement was significant at $p<0.05$ (one-tailed test). The biggest improvements tended to occur for subjects whose scores for the unenhanced speech were between 40 and 60% correct.

In summary, for both types of list and both types of noise, the enhancement process led to significantly improved intelligibility. The improvements were not dramatic, but they occurred consistently.

## DISCUSSION AND CONCLUSIONS

Our initial results are promising, especially since many of the parameters influencing the processing were chosen on the basis of educated guesses. These parameters include: the length of each segment processed; the sampling rate and corresponding resolution of the FFT; the scale of the DOG function relative to the bandwidth of the auditory filter; and the degree of enhancement used. We plan to examine the effects of each of these parameters in order to determine optimum values. We will also examine the patterns of errors made, to determine how successful the enhancement process is for different types of speech sounds. This, in turn, may lead to modifications of the processing. The results so far support the idea that spectral smoothing, followed by the enhancement of spectral contrast, can lead to significant improvements in the intelligibility of speech in noise for the hearing impaired, when both the smoothing and the contrast enhancement are carried out on a frequency scale related to the resolution of the normal ear.

## ACKNOWLEDGEMENT

This work was supported by the Medical Research Council (U.K.).

## REFERENCES

1. Plomp R. Auditory handicap of hearing impairment and the limited benefit of hearing aids. J Acoust Soc Am 1978; 63: 533–49.
2. Glasberg BR, Moore BCJ. Psychoacoustical abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. Scand Audiol 1989; Suppl. 32: 1–25.
3. Tyler RS. Frequency resolution in hearing-impaired listeners. In: Moore BCJ, ed. Frequency selectivity in hearing. London: Academic Press, 1986.
4. Boers PM. Formant enhancement of speech for listeners with sensorineural hearing loss. IPO Annual Progress Report 1980; 15: 21–8.
5. Summerfield AQ, Foster J, Tyler R, Bailey PJ. Influences of formant narrowing and auditory frequency selectivity on identification of place of articulation in stop consonants. Speech Commun 1985; 4: 213–29.
6. Moore BCJ, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am 1983; 74: 750–3.
7. Allen JB. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans Acoust Speech Signal Process 1977; 25: 235–8.
8. Boothroyd A. Developments in speech audiometry. Sound 1968; 2: 3–10.
9. Bench J, Bamford J. Speech hearing tests and the spoken language of hearing-impaired children. London: Academic Press, 1979.

Address for correspondence: B. C. J. Moore, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England