EFFECT OF INDIVIDUALLY TAILORED SPECTRAL CHANGE ENHANCEMENT ON SPEECH INTELLIGIBILITY AND QUALITY FOR HEARING-IMPAIRED LISTENERS

Jing Chen[1,2] *and Brian C.J. Moore*[2]

1. Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, People's Republic of China

2. Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England

ABSTRACT

Most information in speech is carried by spectral changes over time. We determined if enhancing such changes improves the intelligibility of speech in background sounds for hearing-impaired listeners. The values of four parameters controlling the processing were selected for each subject based on a genetic algorithm. The amount of spectral-change enhancement at a given center frequency increased with increasing hearing loss at that center frequency. A control condition was included with no spectral-change enhancement. The intelligibility of speech was measured for sentences in a speech-shaped noise and in multi-talker babble, using two signal-to-masker ratios, specifically chosen for each subject to avoid floor and ceiling effects. The processing led to small but statistically significant improvements in speech intelligibility for both backgrounds and both signal-to-masker ratios.

Index Terms— Spectral change enhancement, speech intelligibility, hearing loss

1. INTRODUCTION

Information in speech is carried by changes in the spectrum over time [1,2]. Such changes may be less audible to hearing-impaired people than to people with normal hearing, because of the reduced frequency selectivity of the former [3,4]. In previous studies [5,6] we developed and evaluated a form of signal processing that enhanced spectral changes. The method, denoted spectral change enhancement (SCE), was based on the fast Fourier transform (FFT): the input signal was segmented, windowed, Fourier transformed, spectrally smoothed, spectral-change enhanced, inverse Fourier transformed, and converted to a running waveform using the overlap-add technique [7]. A spectral-change enhancement function was derived from the pattern of smoothed spectral changes over time by convolution with a difference-of-Gaussians (*DoG*) function. The processing was implemented with four adjustable parameters: *b,* controlling the width of the *DoG* function; ξ and *m,* controlling the effect of preceding frames (determining the amount of smoothing over time); and *S,* controlling the amount of enhancement [5].

Chen *et al*. [5] evaluated the effectiveness of SCE in improving the intelligibility of speech in background sounds for hearing-impaired subjects. The processing improved intelligibility for speech in steady speech-spectrum noise (SSN) but tended to impair intelligibility in a background of two-talker speech (TTS). Large individual differences occurred. Chen *et al.* [6] assessed whether the effectiveness of the SCE was improved when the parameters that controlled the degree and type of enhancement were chosen individually for each subject, using a genetic algorithm (GA) [8], based on subjective preferences for speech clarity. The parameter values selected by the GA varied markedly across subjects. Speech intelligibility was measured for unprocessed stimuli and stimuli processed using the selected parameters, with SSN and TTS maskers and two signal-to-masker ratios (SMRs) for each subject. The intelligibility of speech in the SSN masker at the lower SMR was improved about 14 percentage points by the processing. The overall improvement produced by the processing was significantly larger than the improvement observed by Chen *et al*. [5], suggesting that use of the genetic algorithm was beneficial.

In this study, we assessed the effect of a modified version of the SCE algorithm on the intelligibility and quality of speech in background sounds. The amount of SCE for each frequency varied with the hearing loss at that frequency; the greater the hearing loss, the more SCE was applied, up to a limit. Also, the frequency resolution of the processing (parameter *b*) was allowed to be finer than for our earlier studies. A GA was used to select appropriate parameter values for each subject and each type of background. Details of the GA are given in [6].

2. SIGNAL PROCESSING

The input signal was sampled at 16 kHz and segmented using a 16-ms frame length and 8-ms frame overlap. Each frame was weighted by a 16-ms Hamming window. Then, a 256-point FFT of the windowed segment was calculated, giving 128 magnitude values ($Spec_{org}$) and 128 phase values.

To remove minor irregularities in the spectrum and to preserve major features in the spectrum that would be well represented in a normal auditory system, the magnitude spectrum was transformed to an auditory excitation pattern, using the convolution procedure described by Moore and Glasberg [9]. As a result of this transformation, the original 128 magnitude values were replaced by 128 values denoted *Mag*, which represent a smoothed version of the original spectrum, comparable to the representation in a normal auditory system.

The spectral change across every two adjacent frames was evaluated by expressing the *Mag* values in dB and taking the difference of the *Mag* values for bin *j* in frame *n* and bin *j* in frame *n-1*:

$$R_{j,n}=Mag_{j,n}-Mag_{j,n-1} \quad (1)$$

where if $R_{j,n} > 0$, the magnitude increased from frame *n-1* to frame *n*, and if $R_{j,n} < 0$, the magnitude decreased from frame *n-1* to frame *n*. The magnitude spectrum was modified based on the spectral change values, $R_{j,n}$.

An enhancement function was derived from the spectral change function by convolution with a difference-of-Gaussians (*DoG*) function, which is described by the following equation:

$$DoG(\Delta f)=(1/2\pi)^{1/2}[\exp\{-(2.72\times\Delta f/rb)^2/2\}-(1/2)\exp\{-(2.72\times\Delta f/2rb)^2/2\}] \quad (2)$$

where Δf is the deviation in Hz from the center frequency, *r* is equal to the value of the average equivalent rectangular bandwidth of the auditory filter for ears with normal hearing at that center frequency, ERB_N, as specified by Moore and Glasberg [9], and *b* is a parameter that determines the width of the central lobe of the *DoG* function, measured between the points where the function has zero magnitude. When *b* = 1, the width of the central lobe is 1 ERB_N. The *DoG* function was centered in turn on each frequency in the spectral change function $R_{j,n}$. For a given center frequency of the *DoG* function, the value of the spectral change function was multiplied by the value of the *DoG* function, and the products obtained in this way were summed across frequency. The magnitude value (in dB) of the spectral change function at that frequency was then replaced by that sum. The result of this computation is denoted the enhancement function (*ENF*, in dB). The *ENF* was used to derive modified output spectra, as described below.

The modified magnitude spectrum for a given frame was obtained by adding a gain function to the original magnitude spectrum (both in dB units). It was desired that this gain function was based on *ENF* and was influenced by the smoothed spectrum in a number of preceding frames, with a weight that progressively declined for frames that were earlier in time than the current frame. To achieve this, the gain function for frame *n*, $Gain_n$, was constructed as a weighted average across frames according to the formula:

$$Gain_n=\frac{ENF_n+\xi ENF_{n-1}+\xi^2 ENF_{n-2}+\ldots\xi^m ENF_{n-m}}{1+\xi+\xi^2+\ldots\xi^m} \quad (3)$$

where ξ (≤ 1) is a parameter controlling the relative weighting of earlier frames, and *m* is the number of frames contributing to the weighted average. Then, the value of $Gain_n$ was scaled by multiplying by a factor *S*, which was an adjustable parameter used to control the degree of SCE. Finally, the modified magnitude spectrum of frame *n*, $Spec_{mod}$, was calculated as:

$$Spec_{mod}=Spec_{org}+(S\times Gain_n) \quad (4)$$

Note that the spectrum magnitudes and values of $Gain_n$ are both in dB units. For a given frame of the input signal, the corresponding output signal was created by inverse FFT with the modified magnitude spectrum and the original phases. This was repeated for successive overlapping frames to give the whole processed signal.

The amount of SCE is controlled by parameter *S*, which was constant across frequency in our earlier work. Here, the amount of SCE at each center frequency, *f*, was controlled by parameter *S*(*f*), and increased with increasing hearing loss at *f*, as specified by the formula:

$$S(f)=S'\times 0.03\times HL(f) \quad (5)$$

where *HL(f)* was the subject's audiometric threshold in dB HL at frequency *f* and *S'* was a parameter controlling the overall amount of enhancement. The value of *S'* was selected for each subject using the GA, as described in [6]. The maximum value of *S*(*f*) was limited to the value that would be obtained for a hearing loss of 70 dB. Audiometric thresholds were measured only at 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz. Audiometric thresholds at other frequencies were obtained using cubic-spline interpolation.

A control condition without SCE was also used. The order of testing the SCE and control conditions was counterbalanced across subjects and across the two test sessions.

3. METHOD

3.1. Subjects and compensation for hearing loss

Ten subjects with mild to moderate hearing loss, presumed to be of cochlear origin, were tested. Their ages ranged from 60 to 78 years. The mean audiometric threshold over the range 0.25 to 6 kHz ranged from 29 to 52 dB HL. To compensate for the attenuative effect of the hearing loss, linear amplification according to the Cambridge formula was applied [10] using a finite impulse response filter implemented in Matlab. This is intended to fully restore the audibility of speech with an overall level of 65 dB SPL.

3.2. Equipment

All stimuli were generated using a notebook computer (Intel Core 2 Duo CPU 2.13 GHz, 4.00 GB RAM) with an internal soundcard (Contexant High Definition SmartAudio 221, 16-bit, 16-kHz sampling rate), and presented via Sennheiser HD580 headphones. The overall level of the target plus masker was 65 dB SPL prior to the amplification prescribed by the Cambridge formula. For all tests, subjects were seated in a double-walled sound-attenuating chamber.

3.3. Speech materials

Sentences from the adaptive sentence list (ASL) corpus were used [11] as the target speech for running the GA. Sentences from the IEEE corpus [12] were used for speech intelligibility and quality testing following use of the GA.

Two background sounds were used, speech-shaped noise (SSN) and babble noise (BBN). The SSN had the same long-term average spectral shape as the ASL sentences. The speech used for the BBN background was recorded from male speakers of British English reading naturally from scripts (unrelated to the ASL materials). The BBN was produced by mixing speech from 7 different talkers, each with the same RMS level. The segment of the BBN to be used on a given trial was selected randomly from within a long file. For each trial, the background sound started 500 ms before the target sentence, and finished synchronously with the sentence. For each subject, intelligibility was measured using two SMRs for each background, separated by 4 dB. These are designated low (L) and high (H). The speech and background sound were mixed with the appropriate SMR prior to SCE processing. Note that the SCE processing did not change the overall level or spectral shape of the signals, so any improvements produced by the processing cannot be attributed to better audibility of the processed signals.

3.4. Subjective evaluations

Subjective evaluations of speech pleasantness and clarity were also obtained. The sentences in the background sounds were presented in pairs separated by 300 ms, using the same sentence for a pair. One sentence-in-background was processed and one was unprocessed, and they were randomly assigned to interval 1 and interval 2. The task was to indicate: (1) in which interval is the speech easier to understand? (2) in which interval is the speech more pleasant? They were given the option of responding "same". If the subject could not make a decision, (s)he was allowed to listen to the stimulus again by clicking a "repeat" button. Ratings of the quality of the background sound were not obtained.

4. RESULTS

4.1. Speech intelligibility

The intelligibility of speech in each type of background and for each SMR was measured twice, in two separate test sessions with different sentences. Figure 1 shows mean speech intelligibility scores across subjects. The results for the first and second sessions are displayed in the top and bottom panels, respectively. In each panel, the scores are presented in four groups according to the background type and SMR, with scores for each of the two processing methods (processed versus unprocessed) within each group.

For statistical analysis, to reduce the effects of the bounded percent-correct scores, the scores were transformed into rationalized arcsine units (RAU [13]). Data were averaged across the two test sessions. A three-factor repeated-measures analysis of variance with factors SMR, type of background and processing showed significant effects of SMR [$F(1,9)=14.0$, $p=0.005$], of background type [$F(1,9)=20.1$, $p=0.002$] and of processing [$F(1,9)=135.8$, $p<0.001$]; SCE led to better intelligibility than for the control condition. On average, the SCE led to slightly better intelligibility for both types of background and for both SMRs.

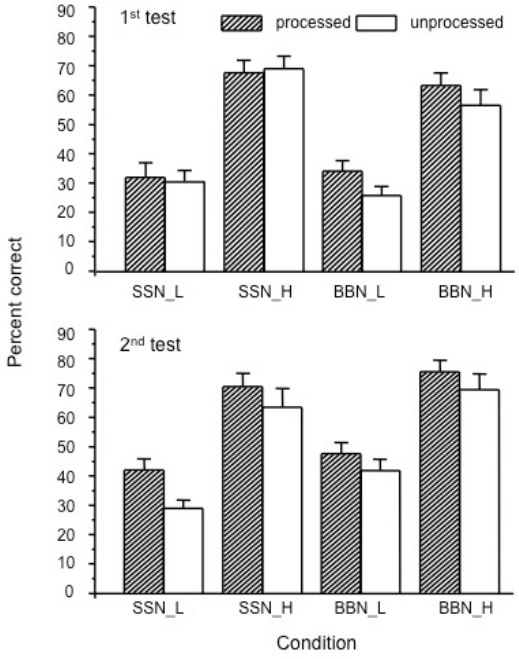

Fig. 1. Mean percent correct identification. Open and shaded bars show scores for unprocessed and SCE processed stimuli, respectively. Error bars indicate the standard error of the mean.

4.2. Subjective evaluations

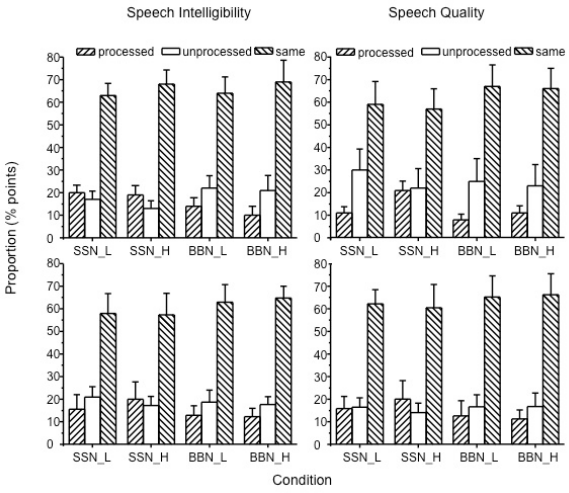

Fig. 2. Mean percentage of selections of each response category for the subjective evaluations. The left and right panels are for speech intelligibility and quality, respectively. The top and bottom panels are for the first and second test sessions, respectively. In each panel, the four groups of bars represent the four test conditions. In each group, the three bars represent the three response categories. Error bars indicate ±1 standard deviation.

Fig. 2 shows the mean percent selections of the three options for subjective speech intelligibility (left panels) and subjective speech quality (right panels) for the first test session (top panels) and the second test session (bottom panels). For each background and both SMRs, the option "they are the same" was selected more than the other two options, indicating that the perceptual effect of the processing was small for these hearing-impaired subjects, whether the question was about speech intelligibility or speech quality. This is consistent with the results of our previous study [6], in which the subjects were asked to compare unprocessed speech and speech processed by SCE (in both cases with a background sound added before processing) and to select "which sentence is more clear", using the same experimental paradigm as here; subjects mostly selected "they are the same". Thus, the SCE led to significant improvements in intelligibility, but did not markedly affect sound quality.

For the first test session, there was a slight trend for the unprocessed speech to be selected more often as higher in quality than the processed speech. However, for the second test session, this trend was not apparent. This may indicate that the subjects became used to the quality of the processed signals with greater exposure, and no longer found the quality to be (slightly) reduced by the processing.

5. DISCUSSION AND CONCLUSIONS

The SCE processing used here differed from that used in our earlier studies, in that the amount of SCE varied across frequency, depending on the amount of hearing loss at each frequency; the greater the hearing loss, the greater was the amount of SCE, up to a specified limit. This meant that the SCE was minimal for frequency regions where the hearing of an individual subject was normal or near-normal, hence limiting the audibility of undesired side effects of the processing. When the SCE was applied uniformly across frequency, as in our earlier work [5,6], the steady background noise had a "gurgling" quality, since random spectro-temporal changes in the noise were magnified by the SCE processing. It appears that making the SCE vary across frequency according to the hearing loss of the subject has beneficial effects, since improvements in intelligibility were found here for both types of background and both SMRs, whereas in our previous studies improvements in intelligibility were found only for the lower SMR and only for the SSN background.

The main conclusions of this study are:

(1) The individually tailored SCE used in this study led to small but significant improvements in the intelligibility of speech in background noise and babble for people with moderate cochlear hearing loss.

(2) The improvements were somewhat greater for the second than for the first test session, which may indicate that the benefit of the processing increases with experience. However, it might also reflect better and/or more reliable performance with the GA on the second test, resulting in better parameter selection.

(3) The results of the subjective evaluations indicated that the effects of the SCE processing on sound quality were relatively small for these hearing-impaired subjects.

6. ACKNOWLEDGMENTS

This work was supported by Starkey, Deafness Research UK, and a Newton International Fellowship from the Royal Society UK. Brian C.J. Moore was supported by the MRC (Grant number G0701870). Thanks to Tao Zhang, Ivo Merks and Kelly Fitz for their helpful discussions and suggestions on this project. The code for the GA was written by Eric Durant and was supplied by Starkey. We thank three reviewers for helpful comments on an earlier version of this paper.

7. REFERENCES

[1] Q. Summerfield, M. P. Haggard, J. Foster and S. Gray, "Perceiving vowels from uniform spectra: phonetic exploration of an auditory after-effect," *Percept. Psychophys.*, vol. 35, pp. 203-213, 1984.

[2] B. C. J. Moore, "Temporal integration and context effects in hearing," *J. Phonetics*, vol. 31, pp. 563-574, 2003.

[3] B. R. Glasberg and B. C. J. Moore, "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," *J. Acoust. Soc. Am.*, vol. 79, pp. 1020-1033, 1986.
[4] B. C. J. Moore, *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues, 2nd Ed.*, Wiley, Chichester, 2007.
[5] J. Chen, T. Baer and B. C. J. Moore, "Effect of enhancement of spectral changes on speech intelligibility for the hearing impaired," *Journal of the Acoustical Society of America*, vol. 131, pp. 2987-2998, 2012.
[6] J. Chen, T. Baer and B. C. J. Moore, "Effect of spectral change enhancement for the hearing impaired based on parameter values selected with a genetic algorithm," *J. Acoust. Soc. Am.*, (in press), 2013.
[7] J. B. Allen, "Short term spectral analysis, synthesis and modification by discrete Fourier transform," *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. 25, pp. 235-238, 1977.
[8] D. Baskent, C. L. Eiler and B. Edwards, "Using genetic algorithms with subjective input from human subjects: implications for fitting hearing aids and cochlear implants," *Ear Hear.*, vol. 28, pp. 370-380, 2007.
[9] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750-753, 1983.
[10] B. C. J. Moore and B. R. Glasberg, "Use of a loudness model for hearing aid fitting. I. Linear hearing aids," *Br. J. Audiol.*, vol. 32, pp. 317-335, 1998.
[11] A. MacLeod and Q. Summerfield, "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use," *Br. J. Audiol.*, vol. 24, pp. 29-43, 1990.
[12] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225-246, 1969.
[13] G. A. Studebaker, "A "rationalized" arcsine transform," *J. Speech Hear. Res.*, vol. 28, pp. 455-462, 1985.