Accepted Manuscript

Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners

Jinqiu Sang, Hongmei Hu, Chengshi Zheng, Guoping Li, Mark E. Lutman, Stefan Bleeck

PII: S0378-5955(15)00159-8

DOI: 10.1016/j.heares.2015.07.019

Reference: HEARES 6989

To appear in: *Hearing Research*

Received Date: 24 December 2014

Revised Date: 20 July 2015

Accepted Date: 23 July 2015

Please cite this article as: Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M.E, Bleeck, S., Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners, *Hearing Research* (2015), doi: 10.1016/j.heares.2015.07.019.

Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners

Jinqiu Sang[a,b], Hongmei Hu[a], Chengshi Zheng[b], Guoping Li[a], Mark E Lutman[a], Stefan Bleeck[a,*]

[a] Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ, UK

[b] Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

[*] Corresponding author. Tel: +4402380596682. Fax: +4402380593190.

E-mail address: s.bleeck@soton.ac.uk

Abstract

Although there are numerous papers describing single-channel noise reduction strategies to improve speech perception in a noisy environment, few studies have comprehensively evaluated the effects of noise reduction algorithms on speech quality for hearing impaired (HI). A model-based sparse coding shrinkage (SCS) algorithm has been developed, and has shown previously (Sang et al., 2014) that it is as competitive as a state-of-the-art Wiener filter approach in speech intelligibility. Here, the analysis is extended to include subjective quality ratings and a method called Interpolated Paired Comparison Rating (IPCR) is adopted to quantitatively link the benefit of speech intelligibility and speech quality.

The subjective quality tests are performed through IPCR to efficiently quantify noise reduction effects on speech quality. Objective measures including frequency-weighted segmental signal-to-noise ratio (fwsegSNR), perceptual evaluation of speech quality (PESQ) and hearing aid speech quality index (HASQI) are adopted to predict the noise reduction effects.

Results show little difference in speech quality between the SCS and the Wiener filter algorithm but a difference in quality rating between the HI and NH listeners. HI listeners generally gave better quality ratings of noise reduction algorithms than NH listeners. However, SCS reduced the noise more efficiently at the cost of higher distortions that were detected by NH but not by the HI.

SCS is a promising candidate for noise reduction algorithms for HI. In general, care needs to be taken when adopting algorithms that were originally developed for NH participants into hearing aid applications. An algorithm that is evaluated negatively with NH might still bring benefits for HI

participants.

Key words

Sparse coding shrinkage; normal hearing; hearing impairment; noise reduction; interpolated paired comparison rating; speech quality

[*]

Abbreviations: BKB, Bamford-Kowal-Bench sentence; CI, cochlear implant; CS-WF, A Wiener filtering approach with cepstral smoothing; fwsegSNR, frequency weighted segmental signal-to-noise ratio; HA, hearing aid; HASQI, hearing aid speech quality index; HI hearing impaired; IPCR, interpolated paired comparison rating; NAL, National Acoustics Laboratory procedure; NH, normal hearing; PESQ, perceptual evaluation of speech quality. SCS, sparse coding shrinkage; SNR, speech-to-noise-ratio; SRT, speech reception threshold; SSN, speech shaped noise.

1. Introduction

For people with mild to severe hearing losses, current advanced hearing aids can help improve speech perception in quiet environments. However, one important reason why hearing-aid users do not use their hearing aids as often as they could is that their devices often do not work well in difficult situations, where for example there is background noise (Alcantara et al., 2003; Dillon, 2001). Hearing-impaired (HI) people typically require a speech-to-noise ratio (SNR) that is at least 3-6 dB higher to achieve the same degree of speech intelligibility (Alcantara et al., 2003; Plomp, 1994) than normal-hearing people. Therefore, noise reduction algorithms in hearing aids are one critical factor to increase hearing aid (HA) uptake and ultimately to improve the quality of life for the HI.

Although microphone arrays have been shown to improve speech perception, their performance is only significant with a large microphone array (Kates et al., 1996; Levitt, 2001; Schum, 2003). Additionally, the microphone array can only be effective when the target speech and interfering sounds are coming from different directions. However, due to the small size of a hearing aid, usually only one or two microphones are placed in a hearing aid. Accordingly, a large microphone array is usually not practical to be placed in a hearing aid. Having two microphones in one hearing aid is effective, because in combination, they allow beam-forming on top of single-channel noise reduction strategies. Currently, most hearing aids are equipped with a combination of single-channel noise reduction algorithms and beam-forming strategies (Widrow et al., 2003) contributing to the overall noise reduction performance. Current commercial devices (e.g. Phonak Naida Q, Starkey 3 series) use wireless transmission between the hearing aids on the two different ears to create more complex beam forming solutions. But still, despite the success of multi-channel approaches, there are many situations that limits the performance of directional microphones and require the use of single-channel strategies, e.g. for the lack of spatial separation between the source and the competition, the effects of reverberation, telephone speech, or hearing aids that are placed entirely in the ear canal. Because of the many limitations of today's single-channel noise reduction algorithms this encourages us to undertake further research.

Previous development of single-channel noise reduction strategies in hearing aids mainly focused on two families of algorithms: spectral subtraction (Alcantara et al., 2003; Dahlquist et al., 2005;

Elberling et al., 1993; Levitt et al., 1993) and Wiener filtering (Levitt et al., 1993). Moreover, most noise reduction algorithms that are currently used in hearing aids, were originally developed to improve speech perception for normal hearing (NH) people and were only later adopted for hearing aid users, thus potentially ignoring specific nonlinear issues of hearing loss like lack of frequency resolution or recruitment. Due to these hearing loss factors in HI listeners, algorithms that are optimal for NH listeners might not be optimal for HI listeners. Also, an algorithm that is not optimal for NH listeners might help HI listeners. Previous studies have found that the same noise reduction algorithm that cannot improve speech intelligibility for NH listeners can significantly improve speech intelligibility for cochlear implant (CI) users (Verschuur et al., 2006; Yang et al., 2005). Whether this phenomenon is similar for the effect of a noise reduction algorithm on speech quality still needs further investigation.

In order for a noise reduction algorithm to be adopted by the users, it should do two things: increasing intelligibility and at the same time keeping or improving the perceived quality. An algorithm might improve speech intelligibility, but introduce distortions to a degree that most users would not want to use it. Many algorithms can do one or the other, but almost all algorithms fail doing both (Hu et al., 2007a; Hu et al., 2008). Therefore it is important to investigate both aspects at the same time.

In our previous paper, a noise reduction strategy based on the principle of sparse coding shrinkage (SCS) was investigated and evaluated in regards to improvement of speech intelligibility in noise with normal hearing and hearing impaired listeners (Sang et al., 2014). The motivation of the strategy is to extract key information from noisy speech to reduce the effects of reduced frequency and temporal resolution. This way, there will be less self-masking and noise-masking of speech components yet essential speech information may be preserved after noise reduction. Of course, this begs the question of how to identify and preserve the key speech information. The approach exploits the principle that the speech signal is highly redundant and information is distributed sparsely in a noisy speech signal. By increasing the sparseness of a noisy speech signal, intelligibility should be improved (Li et al., 2012). The SCS algorithm is based on the assumption that the principal components of clean speech have a super-Gaussian (sparse) distribution. SCS is performed on these principal components. SCS was first proposed by (Hyvärinen, 1999) and first applied to noise reduction in image (Hyvärinen et al., 1998) and later applied to speech enhancement (Hu et al., 2011; Li, 2008; Li et al., 2008; Potamitis et al., 2001; Sang et al., 2011a; Sang et al., 2011b; Zou et al., 2008) in noise. SCS has been shown to be competitive compared to state-of-the-art Wiener filtering in terms of speech intelligibility. However, its performance in speech quality has not been investigated yet. In this paper, speech quality is investigated specifically first using objective (physical) measures and then using subjective quality evaluation with NH and HI listeners.

The most accurate method for evaluating speech quality is through subjective listening tests. Although subjective evaluation of speech enhancement algorithms is often accurate and reliable, it is costly and time consuming. Objective evaluation is an easy and sufficient method to measure an algorithm, as long as the objective measure is in high correlation with subjective tests. Previous research evaluated objective speech quality measures for speech enhancement (Hu et al., 2008). It has been shown that objective measures such as segmental SNR or spectral distortion are not well correlated with speech quality or speech intelligibility (Ma et al., 2009). Other measures are better for this purpose: frequency-weighted segmental SNR (fwsegSNR) and 'perceptual evaluation of speech quality' (PESQ) have both been shown to predict perceived speech quality for NH well (Hu et al., 2008; Ma et al., 2009). Another objective measure 'hearing aid speech quality index' (HASQI) was introduced also to predict speech quality specifically for hearing aid users (Kates et al., 2010). The physical measures fwsegSNR, PESQ and HASQI were adopted in this paper to evaluate the noise reduction algorithms in speech quality.

The most commonly reported subjective techniques of quality evaluation can be broadly classified into two categories: those based on relative preference tasks (Combescure et al., 1982; Hecker et al., 1966; Munson et al., 1962) and those based on assigning a numerical value to the quality of the speech stimuli (Boymans et al., 1999; Harlander et al., 2012; ITU, 2003; Jamieson et al., 1995; Marzinzik, 2000). ITU-T recommendation P.835 (ITU, 2003) was widely used in subjective quality tests (Hu et al., 2006; Hu et al., 2007b). The method instructs the listener to successively attend to and rate the enhanced speech signal using a five-point scale [1=bad, 2=poor, 3=fair, 4=good, 5=excellent]. However, it cannot reflect the difference in signal-to-noise ratio between enhanced speech and unprocessed speech regarding quality impression. A novel method 'Interpolated Paired Comparison Rating' (IPCR) (Dahlquist et al., 2005) was suggested to measure sound quality in terms of subjective SNR gain, which is the difference in SNR between processed and unprocessed speech that give equal subjective sound quality impression. A direct way to measure such an SNR-gain would be to adaptively compare the enhanced stimulus at a fixed SNR against an unprocessed stimulus with a variable SNR. However, this is time-consuming and thus not practical (Dahlquist et al., 2005). Therefore a modified approach was used in this paper to accelerate the measure: IPCR was performed by comparison between processed and unprocessed stimuli at only two different fixed SNRs and the complete function was then extrapolated to find the SNR-gain of subjective equality. The advantage of this modified IPCR approach is that algorithms can be evaluated efficiently and quantitatively by finding subjective equality ratings. The methodology of IPCR was shown to be sensitive enough to detect significant mean differences between enhanced speech and unprocessed speech regarding quality impression. IPCR, with the result of SNR improvement in sound quality, can be compared with speech intelligibility directly on the same (quantitative) scale.

Subjective perceived sound quality can be generally classified in many dimensions including 'preference', 'comfort', 'speech clarity' or 'background noise'. Previous evaluation showed that the dimension of 'preference' could reflect the dimensions of 'comfort' and 'clarity' (Dahlquist et al., 2005). Due to the time restrictions, the dimensions of 'comfort' and 'clarity' are not studied in the present work; however, the performance of the two dimensions can be reflected in the performance of 'preference'. In our study two subjective quality impression dimensions were concentrated on: 'preference' and 'background noise'. This is different from that in Dahlquist et al. (2005) so as to display the ratings of "preference" and "noise loudness" in a more comprehensive comparison. This is because subjective 'preference' is a complex function of several dimensions with different individuals' weights. However, because we are mostly interested in the effect of noise reduction algorithms on speech quality, we aim to separate the selective impact of 'noise loudness' from the overall 'preference'. Comparing between 'preference' and 'noise loudness' thus reveals the weight of 'noise loudness' in the overall subjective rating of 'preference' clearer, although of course both these measures interact. Noise loudness is usually reduced at the price of speech quality, resulting in reduced 'clarity' and 'comfort'. In order to rate 'preference' participants were asked to judge their overall impression of speech quality, including comfort of listening and speech clarity. For 'Noise loudness' we asked participants to judge how much they can perceive the background noise. Participants were asked to rate both quality dimensions on a scale from -10 to10 for each comparison between enhanced stimuli and unprocessed stimuli.

In order to evaluate the results, we compare the SCS algorithm with a Wiener filter as a competitive state-of-the-art noise reduction algorithm (Breithaupt et al., 2008; Gerkmann et al., 2009; Gerkmann et al., 2012) that is frequently used in today's hearing aids. Babble noise and speech shaped noise were chosen as the additive noise. They are challenging for any noise reduction algorithm, because their average long term spectrum is similar to the speech signal.

The structure of this paper is as follows. Firstly, the methods of the two noise reduction algorithms and the quality evaluation are introduced in Section 2. After that, the speech quality performance of the two algorithms are presented and analyzed in Section 3. Factors that affect the performance of noise reduction algorithms in speech quality are discussed in Section 4. Conclusions are presented in Section 5. The objective of the present work is to evaluate the effects of the SCS and CS-WF algorithms for NH and HI listeners in speech quality.

2. Materials and methods

2.1. The sparse coding shrinkage algorithm

Fig. 1 illustrates the flowchart of conducting the sparse coding shrinkage in noisy speech. For details see (Sang et al., 2014). This flowchart is implemented on each divided speech segment with length of N (in our case 64 at a sample rate of 16 kHz). The observed noisy speech is reconstructed into a noisy speech matrix $\mathbf{Z}$. The noisy speech matrix is transformed through $\mathbf{W}^{\mathbf{T}}$ into principal components $\mathbf{Y}$. We assume that clean signals are transformed into a sparse distribution and noise is transformed into a more Gaussian distribution. The shrinkage function $g(\cdot)$ is applied to suppress the noise in noisy components and estimate the clean components. After that, the inverse transform $\mathbf{W}^{-\mathbf{T}}$ and reconstruction is calculated to derive the estimated clean speech signals.

Here, a state-of-the-art noise estimator proposed by (Gerkmann et al., 2012) is adopted to track non-stationary noise. This method estimates the noise power spectral density (NPSD) based on a speech presence probability (SPP), where the a *priori* SNR is a fixed value in estimating the SPP. The amplitude of the noise spectrum is therefore the square root of the noise power spectrum at each frequency bin frame by frame. The phase of the noise spectrum is assumed to be the same as that of the noisy speech spectrum. The noise spectrum can be obtained by multiplying the amplitude of the noise spectrum with the phase of the noisy spectrum. The time-domain noise waveform is accordingly estimated by the inverse FFT of the estimated noise spectrum.

2.2. The Wiener filter as the comparison algorithm

This SCS algorithm was compared with a Wiener filtering approach, of which the code was provided by Timo Gerkmann (Breithaupt et al., 2008; Gerkmann et al., 2009; Gerkmann et al., 2012). One characteristic of this specific approach is that the *a priori* SNR is estimated by the cepstral smoothing method. We thus refer to this approach as ‘CS-WF’ (cepstral smoothing Wiener filter). CS-WF crucially relies on two techniques. First, to estimate the noise power spectral density (NPSD) the speech presence probability (SPP) is calculated. The a *priori* SNR is a fixed value estimating the SPP (Gerkmann et al., 2012). In our own method, the SCS-algorithm, we adopted the same NPSD estimation method. The other technique involved in CS-WF is to estimate the *a priori* SNR using a temporal cepstrum smoothing with bias compensation (Breithaupt et al., 2008; Gerkmann et al., 2009). This algorithm reduces musical noise and suppresses non-stationary noise effectively.

2.3. Stimuli

The speech materials in our experiments were BKB sentences (Bench et al., 1979) recorded by a female talker. The corpus consists of 21 lists with 16 sentences in each list. There are three or four key words in each sentence to make up 50 key words per list. Two noise types were used, babble noise and speech shaped noise. The speech shaped noise was generated by filtering white noise with long term average speech spectrum. The source of babble noise is 100 people speaking in a canteen (http://spib.linse.ufsc.br/noise.html). The room radius is over two meters; therefore, individual voices are slightly audible. The long term average spectra of speech, speech shaped noise and babble noise were shown in Fig. 2. The speech corrupted with noise was further processed with and without noise reduction strategies to produce stimuli conditions that are denoted as ‘noisy’ (no algorithm), ’CS-WF’ (comparison Wiener filter) and ‘SCS’ (sparse coding shrinkage) in this paper.

2.4. Physical Quality Evaluations

2.4.1. Frequency-weighted segmental SNR (fwsegSNR)

The frequency-weighted segmental SNR (fwsegSNR) was computed using the following equation (Hu et al., 2008)

$$fwsegSNR = \frac{10}{M}\sum_{m=0}^{M-1}\left(\left(\sum_{j=1}^{K} W(j,m)\log_{10}\left(X(j,m)^2 / (X(j,m)-\hat{X}(j,m))^2\right)\right) \Big/ \left(\sum_{j=1}^{K} W(j,m)\right)\right)$$

where $W(j,m)$ is the weight placed on the *j*th frequency band, K is the number of bands, M is the total numbe of frames in the signal, $X(j,m)$ is the critical-band magnitude (excitation spectrum) of the clean signal in the *j*th frequency band at the *m*th frame, and $\hat{X}(j,m)$ is the corresponding spectral magnitude of the enhanced signal in the same band. $W(j,m)$ is power of $X(j,m)$, with the power exponent chosen as 0.2 for maximum correlation. 25 bands were used in the calculation. The number of the bands and the shape of weight in each band affect the validation of the measure. They were chosen for maximum correlation between the objective measure and the subjective tests in the study (Hu et al., 2008). The critical-band spectra $X(j,m)$ were obtained by multiplying the FFT magnitude spectra by 25 overlapping Gaussian-shaped windows (Loizou, 2007) spaced in proportion to the ear’s critical bands and summing up the power within each band. Similar to the implementation in (Hu et al., 2008), the excitation spectra were normalized to have an area of unity. This measure has been found to be in good correlation with both subjective quality and intelligibility measures (Hu et al., 2008). The MATLAB code of fwsegSNR was adopted from (Loizou, 2007).

2.4.2. Perceptual Evaluation of Speech Quality (PESQ)

The perceptual evaluation of speech quality (PESQ) (ITU, 2000; Rix et al., 2001) was recommended by the ITU-T for speech quality assessment of handset telephony and narrow-band speech. This measure includes distortions commonly encountered when speech is transmitted via telecommunication networks. The original and degraded signals are first level-equalized to a defined comfortable listening level, and filtered through a standard telephone handset filtering system (300 Hz-3.4 KHz) to simulate handset telephone. The signals are then corrected for time delays and processed through an auditory transform to obtain loudness spectra. The PESQ score is calculated as the difference between these loudness spectra. Resulting PESQ scores are between 1.0 and 4.5, with high values indicating better quality. Although eliminating information above 3.4 kHz may have a differential effect on the noise reduction approaches, PESQ was validated as in good correlation with subjective quality measures (Hu et al., 2008; Rix et al., 2001) for NH listeners. The MATLAB code of PESQ was adopted from (Loizou, 2007).

2.4.3. Hearing Aid Speech Quality Index (HASQI)

Kates and Arehart (2010) proposed the Hearing Aid Speech Quality Index (HASQI) to evaluate speech quality with distortions introduced by hearing aids for both NH and HI listeners. This metric starts with a cochlear model that incorporates aspects of impaired hearing and then extracts signal features related to quality judgments. Two features are used: the effect of noise and nonlinear distortion on speech quality and the effects of linear filtering. The final HASQI index is the multiplicative combination of the nonlinear and the linear effects. The MATLAB code of HASQI was provided by James M. Kates.

2.5. Subjective Quality Evaluation

2.5.1. Procedure of IPCR

Fig. 3 shows the MATLAB GUI used for the paired comparison rating. The principle of this interface is adopted from (Dahlquist et al., 2005). Participants could listen to processed and unprocessed sound by clicking the buttons A or B, but assignment was randomized and they were not aware which one was the processed. Participants were asked to rate their preference between the two stimuli by asking “which one do you prefer?” and “which one has more perceivable background noise?” They could listen to the stimuli as often as they wanted until they reached a final decision. Ratings were recorded by using the slider which was quantified by measurement normalized to a ± 10 scale. The title ‘preference’ or ‘background noise’ was shown to indicate which quality dimension needs to be rated.

Two comparisons for each type of noise with each noise reduction algorithm were used: 1) between processed signal at 5 dB SNR and unprocessed signal at 5 dB SNR (rating value R_0); 2) between processed signal at 5 dB SNR and unprocessed signal at 10 dB SNR (R_5). The ratings of the comparison between processed and unprocessed stimuli were measured for both conditions for each participant. The point of subjective equality was obtained either by linear inter- or extrapolation between these two difference values. Fig. 4 shows an example of this interpolation (Dahlquist et al., 2005). For each participant (index *n*), the SNR improvement G_n is obtained by calculating the point where the line crosses the 0 rating difference. This point indicates how much equivalent quality improvement is achieved by processing the stimulus. So when the result is 5/9 (as in Fig. 4), that indicates that the quality intervention by the algorithm is equivalent to an increase of SNR by 4 dB. This way, the subjective quality rating can be interpreted by an objective equivalent SNR value.

There were a total of eight conditions in this experiment (see Table 1 & Fig. 7): two noise types (SSN, babble); two noise reduction algorithms (CS-WF, SCS) and two SNR conditions (R0 and R5, see above). Two subjective ratings were given in each condition (‘Preference’ and ‘Noise loudness’) and each condition was tested twice. The stimuli in the quality tests were not single sentences, as they are too short, but were very long stimuli so that participants could listen as long as they wanted. The stimuli were generated by concatenating 32 randomly selected BKB sentences. The final presentation of these stimuli was adjusted in level to compensate for the individual hearing thresholds as explained below. The experiment took around half an hour for each participant to complete.

2.5.2. Participants

Nine NH listeners and nine HI listeners with sensorineural hearing loss participated in this experiment. All participants were native English speakers. The NH listeners were recruited from the student population of the University of Southampton and had hearing thresholds at or below 20 dB HL from 250 Hz to 8 kHz, and their ages ranged from 20 to 36. The NH listeners were not informed of the purpose or design of the experiment.

The HI listeners were recruited from volunteers and were all experienced hearing aid users with an age range from 18 to 30. Fig. 5 shows the individual hearing thresholds for the tested ears, showing mostly mild to severe, mostly high frequency hearing losses. All tests were done monaurally on the better ear. The tests were performed via headphones with the hearing aids taken off, and compensation was applied using a linear gain prescription through the NAL-R procedure (Dillon, 2001) according to the individual losses. Table 2 shows the age, tested ear, cause of hearing loss and hearing aid experience of all HI participants. The experiments have been approved by the ethics committee in the University of Southampton.

2.5.3. Equipment

Experiments were performed in a sound-attenuated room at the ISVR Southampton. Stimuli were produced in a MATALB program. Sounds were presented via Sennheiser HDA 200 headphones presented through a Behringer UCA202 sound card and Creek OBH- 21SE headphone amplifier. The presentation levels of speech were kept at 65 dB SPL for NH listeners and were adjusted individually for each HI listener to a subjectively comfortable level.

3. Results

3.1. Results of objective measures

In order to preliminarily understand the results, we first analyzed objective measures fwsegSNR (for noise reduction), PESQ (for quality in NH) and HASQI (for quality in HI). Although HASQI can also reflect the quality for NH listeners, in the present research it only reflects the quality for HI listeners, as hearing threshold parameters corresponding to a moderate hearing loss were assumed in the metric. Using all available 336 BKB sentences (21 lists with 16 sentences in each list) for each measure. Fig. 6 shows all results in a comparison of all conditions. Here, as in the psychophysical experiments, the level of speech was kept constant while the level of noise was varied. No level calibration was necessary, as the physical measures used do not depend on the overall presentation level.

Fig. 6(a) shows the fwsegSNR measure in noisy, CS-WF and SCS conditions in speech shaped noise (left half panel) and babble noise (right half panel) under 0, 5, 10 dB input SNRs. In most conditions, SCS provided the highest noise reduction in speech shaped noise, and CS-WF provided the highest noise reduction in babble noise.

Fig. 6(b) shows the equivalent results for PESQ scores. CS-WF shows best speech quality in both speech shaped noise and babble noise. This indicates that although SCS achieved greater noise reduction (Fig. 6(a)), it introduces distortions, and this is reflected by the reduced PESQ scores.

Fig. 6(c) shows the equivalent results of the HASQI measures. This metric incorporates the parameters of an average hearing loss of our participants, reflecting a moderate hearing loss. Similar to fwsegSNR, SCS and CS-WF show the highest quality in speech shaped noise and babble noise respectively.

Statistical analysis through a three-way repeated ANOVA is done in each plot of Fig. 6. The effects of algorithm, noise condition and input are all significant ($p<0.05$). Each objective measure represents a model that simulates the average performance of subjects, and the effects of factors with the same model tend to be in accordance. As the models in the objective measures are different, the results of the three objective measures are not consistent. More reliable performance of algorithms can be investigated through subjective tests.

3.2. Results of subjective speech quality experiments

Fig. 7 shows the median values of Paired Comparison Ratings for the difference between processed and unprocessed speech for two rating categories (‘preference’ and ‘noise loudness’). The ratings are always presented in a way that a positive value indicates preference towards (or less perceived noise of) the processed speech (against the unprocessed speech). The filled bars in Fig. 7 represent the median ratings of the differences (processed minus unprocessed) for each noise condition, error bars are the inter-quartile range. Ratings for the various input SNRs and noise types are plotted in separate bars. Higher bars indicate preference (or higher rating) of the processed signal compared to the unprocessed signal. The results are shown in separate plots for the NH group and HI. All four plots in Fig. 7 show positive bars (except one bar in plot (a)) and thus indicate generally preference of processed speech against unprocessed speech. Comparison between plot (a) and plot (b) indicates that NH listeners do not prefer the quality of noise reduction algorithms as much as HI listeners do. Plot (a) shows that NH participants prefer the quality of CS-WF compared to SCS. Plot (c) shows that NH listeners can clearly perceive the benefits of noise reduction algorithms as indicated by the reduced 'noise loudness'. The difference between 'preference' (a) and 'noise loudness' (c) indicates that NH listeners have based their judgments not only on 'noise loudness' but also on other perceptual dimensions (e.g. 'clarity', 'comfort', etc.). Plots (b) and (d) indicate that HI listeners get obvious benefits from both noise reduction algorithms in both categories. Comparing (a) and (b) shows that HI participants are less sensitive to speech distortion than NH participants when asking for 'preference'. Furthermore, they are also less sensitive to 'noise loudness', as shown by comparing (c) and (d). The results in plot (b) and plot (d) are comparable with the results in Dahlquist *et al.* (2005) where HI participants gave similar range of quality ratings in 'preference' and 'noise loudness' when tested using a nonlinear spectral subtraction algorithm.

The effects of the algorithm, noise type and SNR (5/5 or 5/10 as explained before) in each plot of Fig. 7 were analyzed through a two-way repeated ANOVA. In Fig. 7(a), for NH subjects assessing 'preference', neither the effect of algorithm nor the effect of noise type is not significant [$F(1,8)=1.1$, $p>0.05$, and $F(1,8)= 1.0$, $p>0.05$, respectively], but the effect of SNR is significant [$F(1,8)=24.4$, $p<0.05$]. In Fig. 7(b), for HI subjects assessing 'preference', the effects of algorithm, noise type and SNR are all significant [$F(1,8)=6.0$, $p<0.05$, $F(1,8)=16.8$, $p<0.05$, and $F(1,8)=48.3$, $p<0.05$, respectively]. This indicates that HI subjects may perceive SCS better than CS-WF. In Fig. 7(c), for NH subjects assessing 'noise loudness', neither the effect of algorithm nor the effect of noise type is not significant [$F(1,8)=0.68$, $p>0.05$, and $F(1,8)=4.1$, $p>0.05$, respectively], but the effect of SNR is significant [$F(1,8)=28.5$, $p<0.05$]. In Fig. 7(d), for HI subjects assessing 'noise loudness', the effect of algorithm is not significant [$F(1,8)=2.8$, $p>0.05$], but the effects of noise type and SNR are significant

[$F(1,8)=6.4$, $p<0.05$, and $F(1,8)=93.9$, $p<0.05$, respectively]. In all plots of Fig. 7 except Fig. 7(b), there is no significant effect of algorithm, indicating that there is no obvious advantage of SCS compared to CS-WF in speech quality.

Median SNR improvements across participants in the normal hearing group and hearing impaired group are presented in Table 3. The individual SNR improvement measures were obtained by IPCR. Each SNR improvement with a positive value indicates that noise reduction algorithms improved speech quality for NH or HI listeners. A Wilcoxon matched-pairs signed ranks test was used to test if the SNR improvements were significantly larger than 0 dB or 5 dB (at a 5%-level; * indicates SNR-improvement >0 dB, ** indicates SNR-improvement >5 dB). The improvements were limited to +10 dB, because single extrapolated values may become incidentally very large (+10 dB was selected as this was assumed to beyond the maximum expected benefit).

Table 3 demonstrates clearly how NH and HI listeners benefit differently from the noise reduction algorithms: For both 'preference' and 'noise loudness', HI values are all larger than NH values, indicating that in all conditions hearing impaired listeners get larger equivalent SNR improvements. This shows that noise reduction algorithms are more beneficial in term of speech quality for HI listeners than for NH listeners.

Interestingly, for NH the SNR improvement of the 'noise loudness' rating is always larger than that of the 'preference' under same conditions. This indicates that although NH listeners notice a reduction in 'noise loudness', this only partially contributes to the final judgment of overall quality 'preference'. This shows that, although noise reduction algorithms reduce noise, they do that at the price of introducing distortion. NH listeners are more easily affected by speech distortion than HI listeners.

Table 4 shows the comparison between median subjective SNR improvement in 'preference' across NH participants and the objective quality measure fwsegSNR improvement. FwsegSNR improvement indicates fwsegSNR of processed speech minus fwsegSNR of unprocessed speech assuming an input SNR of 5 dB. The table shows the SCS algorithm achieved higher 'preference' in speech shaped noise but lower in babble noise. The objective measure fwsegSNR confirms this result, as the correlation between the subjective SNR improvement in 'prefernece' for NH and the objective fwsegSNR improvement is high (r=0.96). Quantitative correlation analysis is also done between subjective measures and other objective measures. The correlation coefficients are high between PESQ and subjective SNR improvement in ‘preference’ for NH (r=0.54), between PESQ and subjective SNR improvement in ‘noise loudness’ (r=0.8). The correlation coefficients are low between HASQI and subjective SNR improvement in ‘preference’ for HI (r=0.1), between HASQI and subjective SNR improvement in ‘noise loudness’ for HI (r=0.05). However, only the condition of 5 dB input SNR is considered in comparison. Whether the objective measures are correlated with subjective measures at other input SNRs are not investigated. The measures disclose different aspects of the processing and it is important to include a range of measures. The purpose is not to carry out research on the measures themselves. By giving a range of measures, the results can be compared with other studies that have used those measures.

3.3. Speech intelligibility versus speech quality

The interesting result of the IPCR method is that it allows comparing sound quality and speech intelligibility directly on the same (quantitative) scale. The results of the speech intelligibility measures that are used here were previously published in (Sang et al., 2014) and were measured with the same participants (by the same authors). The speech intelligibility studies were conducted before the speech quality tests. There was no negative learning effect as the two tests evaluated two different dimensions. Even if there exists any learning effect, familiarity with the speech materials may help judge the quality of the speech. The speech intelligibility test was performed with BKB sentences using the same noises and same noise reduction algorithms as shown here. We used a three-up-one-down adaptive procedure as described in (Sang et al., 2014) to find the speech reception threshold (SRT in dB) required for 79.4% correct recognition. Fig. 8 provides an intuitive visualization of all results allowing a quantitative comparison between speech quality and intelligibility. This allows answering the following question quantitatively: how much benefit does the noise reduction algorithms have on speech quality and intelligibility?

Fig. 8 show the individual results of all participants in each noise type (columns) and algorithms (marker type) for NH (left) and HI (right). Each plot shows the benefit in speech recognition (x-axis) vs. the benefit in 'preference' (y-axis) in dB. The improvement in 'preference' was derived from the IPCR method as explained above, the improvement in speech recognition was calculated as the SRT of the unprocessed speech minus the SRT of the processed speech.

Improvements were generally smaller for speech intelligibility than for quality. This is the case for all four plots, which means that both noise reduction algorithms improve speech quality more than speech intelligibility. However, this does not necessarily mean that these two effects do not contribute to the individuals' overall benefits with the same weight. Results from NH participants in speech shaped noise (shown in plot a) show that the improvements for speech intelligibility are clustered around 0 dB indicating that there are no large differences between processed speech and unprocessed speech. For NH participants listening to speech in babble noise (shown in plot c), the improvements in speech recognition are almost all in the left half of the plot (less than 0 dB), which reflects a worse intelligibility in the processed speech. For HI participants listening in speech shaped noise (b) and babble noise (d), almost all SNR improvements in speech intelligibility are in the right half of the plot (above 0 dB). This is further evidence that noise reduction algorithms benefit intelligibility for most HI participants but not for NH participants. By inspecting the speech intelligibility improvements with respect to the horizontal axis, the speech recognition gains of CS-WF and SCS scatter in a similar range in each plot which further suggests little difference between the algorithms in terms of intelligibility. For NH participants, there are several negative 'preference' improvements for SCS, indicating quality reduction from the noise reduction algorithms. In (a) and (c) 'preference' improvements are generally less for SCS than for CS-WF. This means that for NH, SCS reduces noise more effectively, but introduces more speech distortion.

For HI participants evaluating speech quality in (b) and (d), all 'preference' improvements are positive and most of them are near 10 dB which evidencing again that both noise reduction algorithms improve speech quality more for HI participants than for NH participants.

Fig. 8 suggests that there is the potential for a modest speech intelligibility improvement for HI listeners without any significant cost in sound quality. Often, more aggressive signal processing strategies such as SCS are rejected either in the design process or the clinical use stage because of judgments made by normal hearing engineers or clinicians. These data call into question those practices.

4. Discussion

4.1. Choice of noise conditions

It is important to discuss the appropriate choice of noise type as a methodological issue. Noise that contains mainly low-frequency components is easier to remove from speech compared to noise that shows a similar frequency spectrum as speech (Dillon et al., 1993; Elberling et al., 1993). Dillon and Lovegrove (1993) found that the benefit of previous single-channel noise reduction systems in terms of speech intelligibility was small and the amount of improvement was greatest when the noise spectrum was weighted towards low frequencies. White noise and pink noise are also easy to reduce as they show a different pattern of spectrum from speech. To make the situation of speech in noise more realistic and difficult, speech shaped noise and babble noise were chosen in our study, both of which show similar average spectra as the speech.

However, it should be pointed that even clinical babble fails to truly capture the task of listening in realistic environments. Competing signals with some level of linguistic content in the competition come closer to replicating the true problems of those with HI, but those competing signals are likely more difficult for the signal processing techniques considered in this paper to manage.

4.2. Difference between NH and HI participants in benefits from algorithms

The speech quality tests showed that NH and HI participants noticed the noise reduction effects of both algorithms; however, HI participants had higher ‘preference’ improvements than NH participants in almost all cases. It appears that HI listeners are less sensitive to speech distortion but more sensitive to noise loudness. Noise reduction strategies are usually developed to reduce noise at the price of introducing speech distortion, which might be easily perceived by NH listeners but not by HI listeners. That may be one reason why noise reduction strategies are more beneficial to HI listeners than to NH listeners in terms of speech quality. This is in accordance with (Schijndel et al., 2001) where it was shown that HI participants were less sensitive to spectral distortion; when speech was distorted speech intelligibility degrades significantly in NH participants but not in HI participants.

4.3. Comparison between SCS and CS-WF

On the whole, there was no large difference in performance between SCS and CS-WF either within NH participants or HI participants. CS-WF was processed in the frequency domain while SCS was processed in the eigenvalue domain. CS-WF uses cepstrum smoothing technique when estimating a priori SNR. SCS is a model-based approach which assumes super-Gaussian distribution of speech, while CS-WF does not assume data distribution but adaptively filtering the corrupted speech in the frequency domain. The clean speech is in a complex super-Gaussian distribution, and the assumed speech distribution in SCS could not match the clean speech distribution very well. That is the reason why SCS did not show obvious advantage over CS-WF on speech quality.

However, as SCS reduced more noise, more distortion was detected by NH than HI. SCS presents sparse stimuli with a larger degree of noise reduction, which is more acceptable to HI participants who are less sensitive to speech distortion and more sensitive to noise level due to the hearing loss factors. Thus one needs to be careful when adopting algorithms that were originally developed for NH participants into hearing aid or cochlear implant applications. An algorithm that is evaluated negatively with NH might still bring benefits for HI participants in speech intelligibility and quality.

4.4. Acclimatization effects

Acclimatization is the process when a listener is adjusting to a gradual change in environment, allowing maintaining performance across a range of environmental conditions. Acclimatization effects are important for the individual performance with noise reduction algorithms. The speech quality test here recruited the same participants and used the same speech materials as the speech recognition test before. Therefore, the speech recognition test could be regarded as training practice for the speech quality test. Therefore, there might be some learning effect in the speech quality test, which can help participants give a fair quality comparison of different algorithms.

5. Conclusions

This work focuses on the benefits of single-channel noise reduction algorithms on speech quality in HI listeners. The motivation was to evaluate whether our newly developed SCS algorithm benefits HI listeners more than a state-of-the-art competitive noise reduction algorithm (CS-WF) in terms of speech quality. The experiments show that there was no particular benefit of SCS compared to CS-WF on speech quality.

The objective measures fwsegSNR and PESQ are in high correlation with subjective measures, while another objective measure HASQI is not correlated with subjective measures in present research. Both algorithms benefited all listeners in terms of speech quality. However, HI listeners got more benefits from the noise reduction algorithms in speech quality than NH listeners. Both algorithms have the potential to provide a modest intelligibility improvement for HI listeners without any significant cost in sound quality. We conclude that algorithms that are evaluated negatively with NH listeners might still benefit HI listeners.

Acknowledgements

We thank Aapo Hyvarinen, Patrik Hoyer and Xin Zou for their advice in sparse coding shrinkage. We thank Timo Gerkmann for providing CS-WF code. We thank David Simpson, James M. Kates and Kathryn Hoberg Arehart for their advice in NAL-R compensation. We also thank all the participants. This work was supported by the European Commission within the ITN AUDIS (Grant No. PITN-GA-2008-214699). This work was partially supported by the National Natural Science Foundation of China (Grant No. 61302168).

ACCEPTED MANUSCRIPT

References

Alcantara, J.I., Moore, B.C., Kuhnel, V., Launer, S. 2003. Evaluation of the noise reduction system in a commercial digital hearing aids. Int. J. Audiol. 42, 34-42.

Bench, J., Kowal, A., Bamford, J. 1979. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. Brit. J. Audiol. 13, 102-112.

Boymans, M., Schoneveld, P., Verschuure, J. 1999. Clinical evaluation of a full-digital in-the-ear hearing instrument. Audiology 38, 99-108.

Breithaupt, C., Gerkmann, T., Martin, R. 2008. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing, IEEE ICASSP 2008. . pp. 4897-4900.

Combescure, P., Le Guyader, A., Gilloire, A. 1982. Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings, IEEE ICASSP '82., Vol. 7. pp. 988-991.

Dahlquist, M., Lutman, M.E., Wood, S., Leijon, A. 2005. Methodology for quantifying perceptual effects from noise suppression systems. Int. J. Audiol. 44, 721-732.

Dillon, H. 2001. Hearing Aids, Thieme, New York.

Dillon, H., Lovegrove, R. 1993. Single-microphone noise reduction systems for hearing aids: A review and evaluation. In: Studebaker, G.A., Hochberg, I., (Eds.), Acoustical factors affecting hearing aid performance. Allyn and Bacon, Boston. pp. 353-372.

Elberling, C., Ludvigsen, C., Keidser, G. 1993. The design and testing of a noise reduction algorithm based on spectral subtraction. Scand. Audiol. Suppl. 1993, 39-49.

Gerkmann, T., Martin, R. 2009. On the Statistics of Spectral Amplitudes After Variance Reduction by Temporal Cepstrum Smoothing and Cepstral Nulling. Signal Processing, IEEE Transactions on 57, 4165-4174.

Gerkmann, T., Hendriks, R.C. 2012. Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. Audio, Speech, and Language Processing, IEEE Transactions on 20, 1383-1393.

Harlander, N., Rosenkranz, T., Hohmann, V. 2012. Evaluation of model-based versus non-parametric monaural noise-reduction approaches for hearing aids. Int. J. Audiol., 1-13.

Hecker, M.H.L., Williams, C.E. 1966. Choice of Reference Conditions for Speech Preference Tests. J. Acoust. Soc. Am. 39, 946-952.

Hu, H., Li, G., Chen, L., Sang, J., Wang, S., Lutman, M.E., Bleeck, S. 2011. Enhanced sparse speech processing strategy for cochlear implants, Eusipco, Barcelona, Spain.

Hu, Y., Loizou, P.C. 2006. Subjective Comparison of Speech Enhancement Algorithms, Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, Vol. 1. pp. I-I.

Hu, Y., Loizou, P.C. 2007a. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122, 1777-1786.

Hu, Y., Loizou, P. 2007b. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49, 588-601.

Hu, Y., Loizou, P.C. 2008. Evaluation of objective quality measures for speech enhancement. IEEE T. Speech Audi. P. 16, 229-238.

Hyvärinen, A. 1999. Sparse code shrinkage: Denoising of nonGaussian data by maximum likelihood estimation. Neural Comput. 11, 1739-1768.

Hyvärinen, A., Hoyer, P., Oja, E. 1998. Sparse code shrinkage for image denoising, Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence., Vol. 2. pp. 859-864.

ITU. 2000. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation. pp. 862.

ITU. 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. ITU-T Recommendation. pp. 835.

Jamieson, D.G., Brennan, R.L., Cornelisse, L.E. 1995. Evaluation of a speech enhancement strategy with normal-hearing and hearing impaired listeners. Ear. Hear. 16, 274-286.

Kates, J., Arehart, K.H. 2010. The hearing-aid speech quality index (HASQI). J. Audio Eng. Soc. 58, 363-381.

Kates, J.M., Weiss, M.R. 1996. A comparison of hearing-aid array-processing techniques. J. Acoust. Soc. Am. 99, 3138-3148.

Levitt, H. 2001. Noise reduction in hearing aids: A review. J. Rehabil. Res. Dev. 38, 111-121.

Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., Weiss, M. 1993. Signal processing for hearing impairment. Scand. Audiol. Suppl. 38, 7-19.

Li, G. 2008. Speech perception in a sparse domain. PhD thesis, Institute of Sound and Vibration, University of Southampton.

Li, G., Lutman, M.E. 2008. Sparse Stimuli For Cochlear Implants, Proc. EUSIPCO, Lausanne, Switzerland.

Li, G., Lutman, M.E., Wang, S., Bleeck, S. 2012. Relationship between speech recognition in noise and sparseness. Int. J. Audiol. 51, 75-82.

Loizou, P.C. 2007. Speech Enhancement: Theory and Practive, CRC, Talor & Francis.

Ma, J., Hu, Y., Loizou, P. 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125, 3387-3405.

Marzinzik, M. 2000. Noise reduction schemes for digital hearing aids and their use for the hearing-impaired, University of Oldenburg.

Munson, W.A., Karlin, J.E. 1962. Isopreference Method for Evaluating Speech-Transmission Circuits. J. Acoust. Soc. Am. 34, 762-774.

Plomp, R. 1994. Noise, amplification, and compression: Considerations of three main issues in hearing aid design. Ear Hearing 15, 2-12.

Potamitis, I., Fakotakis, N., Kokkinakis, G. 2001. Speech enhancement using the sparse code shrinkage technique, IEEE ICASSP '01. pp. 621-624.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, IEEE ICASSP '01, Vol. 2. pp. 749-752.

Sang, J., Li, G., Hu, H., Lutman, M.E., Bleeck, S. 2011a. Supervised sparse coding strategy in cochlear implants, Interspeech, Florence, Italy.

Sang, J., Hu, H., Li, G., Lutman, M.E., Bleeck, S. 2011b. Supervised sparse coding strategy in hearing aids, IEEE ICCT 2011 China.

Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M.E., Bleeck, S. 2014. Evaluation of the sparse coding shrinkage noise reduction algorithm in normal hearing and hearing impaired listeners. Hear. Res. 310, 36-47.

Schijndel, N.H.v., Houtgast, T., Festen, J.M. 2001. Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 110, 529-542.

Schum, D.J. 2003. Noise-reduction circuitry in hearing aids, II: Goals and current strategies. Hearing Journal 56, 32-41.

Verschuur, C., Lutman, M.E., Wahat, N.H.A. 2006. Evaluation of a non-linear spectral subtraction noise suppression scheme in cochlear implant users. Cochlear Implants Int. 7, 188-193.

Widrow, B., Luo, F. 2003. Microphone arrays for hearing aids: An overview. Speech Commun. 39, 139-146.

Yang, L., Fu, Q. 2005. Spectral subtraction-based speech enhancement for cochlear implant

patients in background noise. J. Acoust. Soc. Am. 117, 1001-1004.
Zou, X., Jancovic, P., Ju, L., Kokuer, M. 2008. Speech Signal Enhancement Based on MAP Algorithm in the ICA Space. IEEE T. Signal Proces 56, 1812-1820.

Captions

Table 1

Combinations of speech-to-noise ratios used for the various types.

Table 2

Age, tested ear, cause of hearing loss and hearing aid experience of the listeners with hearing losses. All of them are bilateral hearing impaired.

Table 3

Median values for SNR-improvement (in dB) for the rating categories ‘preference’ and ‘noise loudness’. ** indicates significant SNR-improvement >5 dB, * indicates significant SNR-improvement >0 dB.

Table 4

Objective and subjective noise reduction effects (dB) for babble noise and speech shaped noise. Improved frequency-weighted segmental SNR (fwsegSNR) and subjectively estimated with IPCR method for ‘preference’ criterion from normal hearing group.

Fig. 1. Flowchart of sparse coding shrinkage algorithm in noisy speech.

Fig. 2. Long term average spectra of speech, speech shaped noise and babble noise.

Fig. 3. MATLAB GUI used for paired comparison rating of speech quality. Participants were asked to rate ‘preference’ (in this case) and ‘noise loudness’ comparing two stimuli (A and B). The buttons “A” and “B” allow unlimited repetition of the stimuli. Participants indicate their rating by adjusting the slider continuously between −10 and 10.hart of simultaneous diagonalization of the estimated speech and noise covariance matrices.

Fig. 4. The method used to estimate the point of subjective equality (PSE) (filled square) from two paired comparison ratings (filled circles) , calculated by linear interpolation or extrapolation(IPCR method (Dahlquist et al., 2005)). In this example, the pair of SNRs for subjective equality is interpolated to 5/9 dB for processed/unprocessed stimuli, indicating an equivalent SNR improvement of +/4 dB. See text for more information.

Fig. 5. Audiograms showing the individual hearing thresholds for the aided ears of HI subjects (N=9).

Fig. 6. Results of objective measures of fwsegSNR, PESQ and HASQI. A more positive value corresponds to better performance in each measure.

Fig. 7. Subjective ratings from paired comparison rating tests for the two sound quality dimensions (upper row: ‘Preference’; lower row: ‘Noise loudness’) for both noise types and both noise reduction algorithms. Left: NH; right: HI. The bars show the median scores of the difference between processed and unprocessed signals (error bars: inter-quartile range). SSN: speech shaped noise. Labels, such as 5/10, indicate (SNR processed) / (SNR unprocessed) in dB. Larger values indicate greater preference for processed speech.

Fig. 8. Scatter plots showing individual 'preference' improvements vs. speech recognition gains in both noises (top vs. bottom) with both noise reduction algorithms (diamonds vs. triangles) with NH and HI listeners (left vs. right).

ACCEPTED MANUSCRIPT

Table 1

Combinations of conditions used. There were a total of eight conditions in this experiment: two noise types (SSN, babble); two noise reduction algorithms (CS-WF, SCS) and two SNR conditions.

Noise type	*Noise reduction method*	*Speech-to-noise ratio (dB) for processed/unprocessed item*
Speech shaped noise	CS-WF	5/5
Babble noise	SCS	5/10

Table 2

Age, tested ear, cause of hearing loss and hearing aid experience of the listeners with hearing losses. All of them are bilateral hearing impaired.

Listener	Age	Gender	Ear	Cause of hearing loss	Hearing aid experience
HI1	20	F	R	meningitis at 2 years old	16 years
HI2	31	F	R	Congenital	31 years
HI3	22	F	R	Congenital	20 years
HI4	18	F	R	Congenital	14 years
HI5	21	F	R	Congenital	18 years
HI6	20	F	R	Tinnitus, noise exposure	4 years
HI7	22	M	L	Congenital	18 years
HI8	20	F	R	Congenital	19 years
HI9	22	M	R	congenital, hereditary	6 years

Table 3

Median values for SNR-improvement (in dB) for the rating categories “preference” and “noise loudness”. ** indicates significant SNR-improvement >5 dB, * indicates significant SNR-improvement >0 dB.

Rating Category	Hearing Level	CS-WF SSN	CS-WF Babble	SCS SSN	SCS Babble
Preference (dB)	NH	5.9**	7.3**	8.8**	3.1*
	HI	10**	9.9**	10**	10**
Noise Loudness (dB)	NH	9.1**	9.7**	10**	10**
	HI	9.8**	10**	10**	10**

Table 4

Objective and subjective noise reduction effects (dB) for babble noise and speech shaped noise. Improved frequency-weighted segmental SNR (fwsegSNR) and subjectively estimated with IPCR method for 'preference' criterion from normal hearing group.

Noise reduction effect	SSN CS-WF	SSN SCS	Babble CS-WF	Babble SCS
Subjective (dB)	5.9	8.8	7.3	3.1
fwsegSNR (dB)	1.0	1.5	1.0	0.5

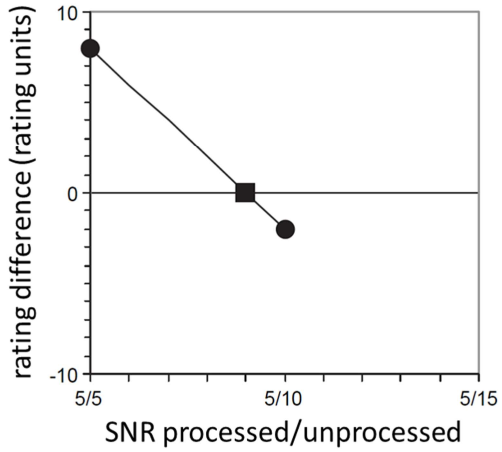

Fig. 4. The method used to estimate the point of subjective equality (PSE) (filled square) from two paired comparison ratings (filled circles) , calculated by linear interpolation or extrapolation(IPCR method (Dahlquist et al., 2005)). In this example, the pair of SNRs for subjective equality is interpolated to 5/9 dB for processed/unprocessed stimuli, indicating an equivalent SNR improvement of +/4 dB. See text for more information.

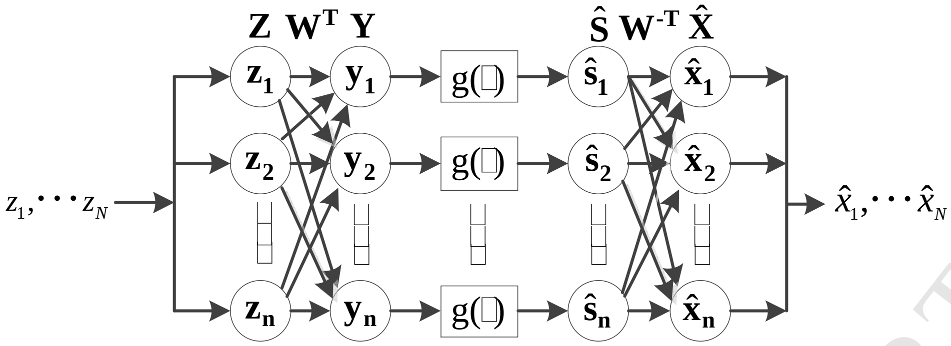

Fig. 1. Flowchart of sparse coding shrinkage algorithm in noisy speech.

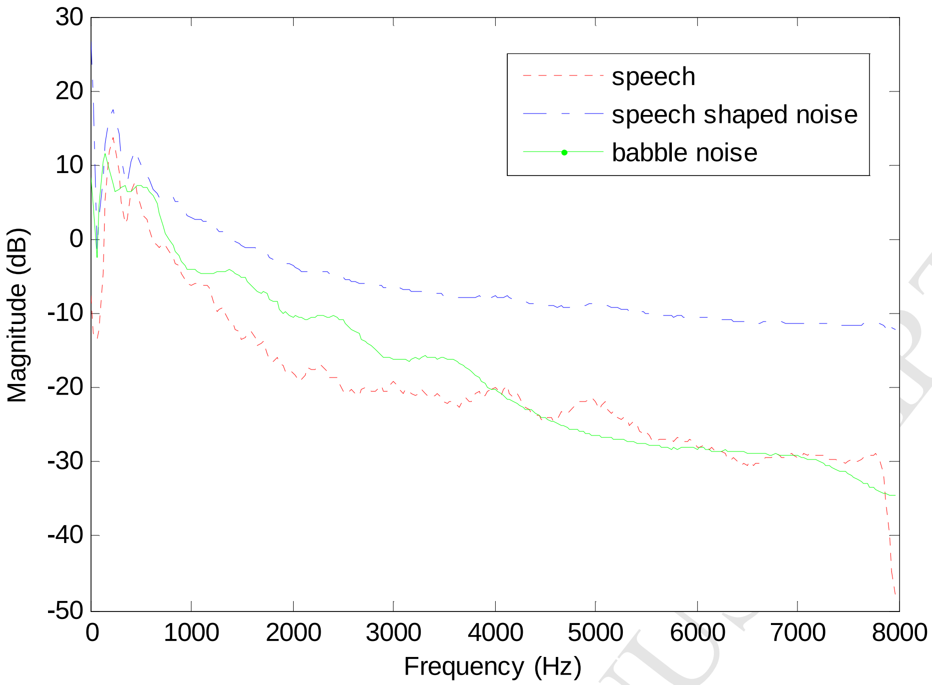

Fig. 2. Long term average spectra of speech, speech shaped noise and babble noise.

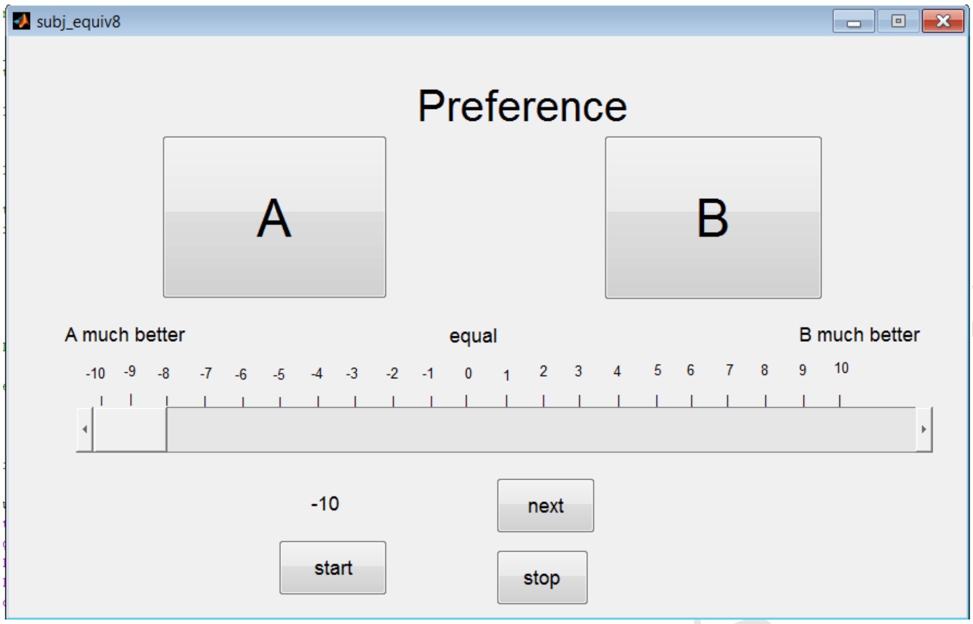

Fig. 3. MATLAB GUI used for paired comparison rating of speech quality. Participants were asked to rate ‘preference’ (in this case) and ‘noise loudness’ comparing two stimuli (A and B). The buttons “A” and “B” allow unlimited repetition of the stimuli. Participants indicate their rating by adjusting the slider continuously between −10 and 10.

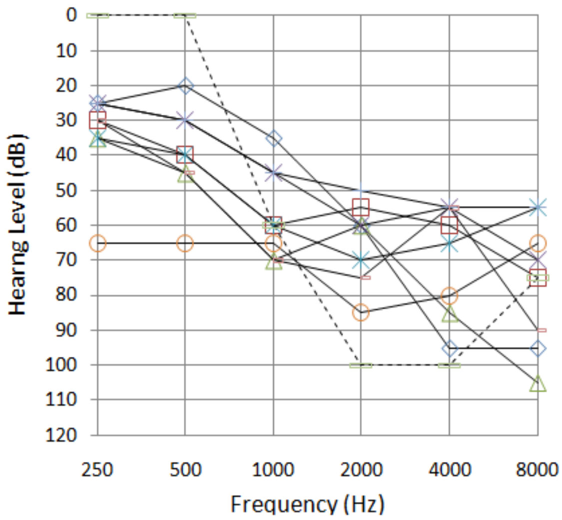

Fig. 5. Audiograms showing the individual hearing thresholds for the aided ears of HI subjects (N=9).

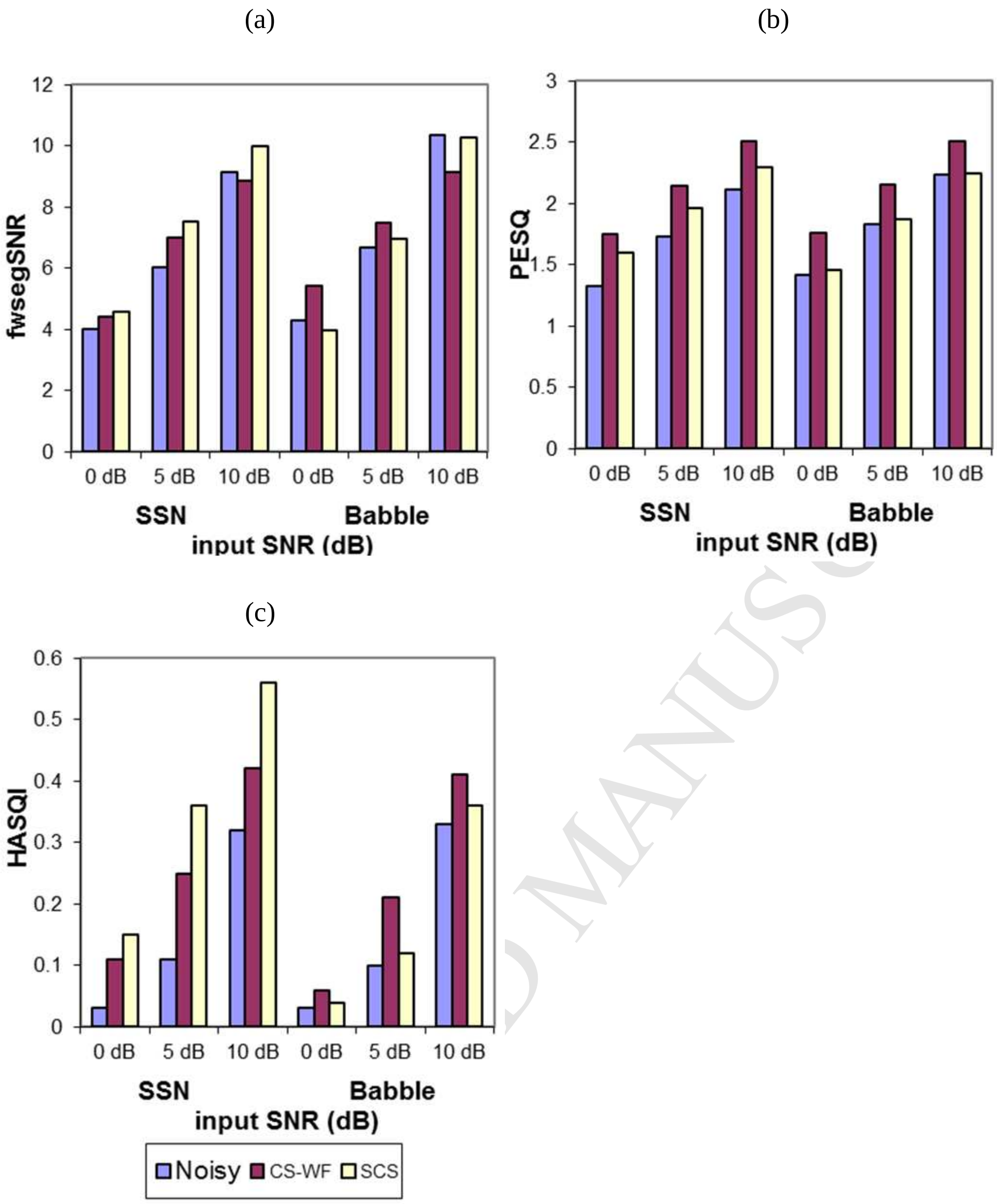

Fig. 6. Results of objective measures of fwsegSNR, PESQ and HASQI. A more positive value corresponds to better performance in each measure.

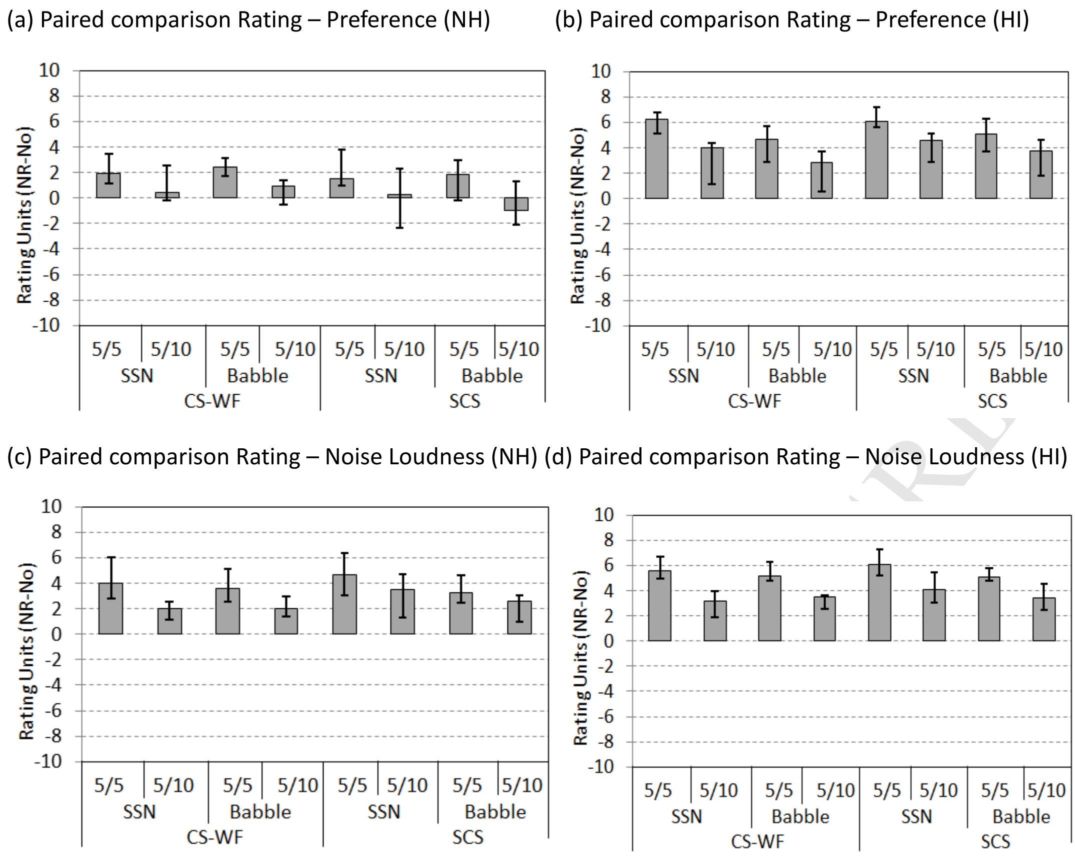

Fig. 7. Subjective ratings from paired comparison rating tests for the two sound quality dimensions (upper row: 'Preference'; lower row: 'Noise loudness') for both noise types and both noise reduction algorithms. Left: NH; right: HI. The bars show the median scores of the difference between processed and unprocessed signals (error bars: inter-quartile range). SSN: speech shaped noise. Labels, such as 5/10, indicate (SNR processed) / (SNR unprocessed) in dB. Larger values indicate greater preference for processed speech.

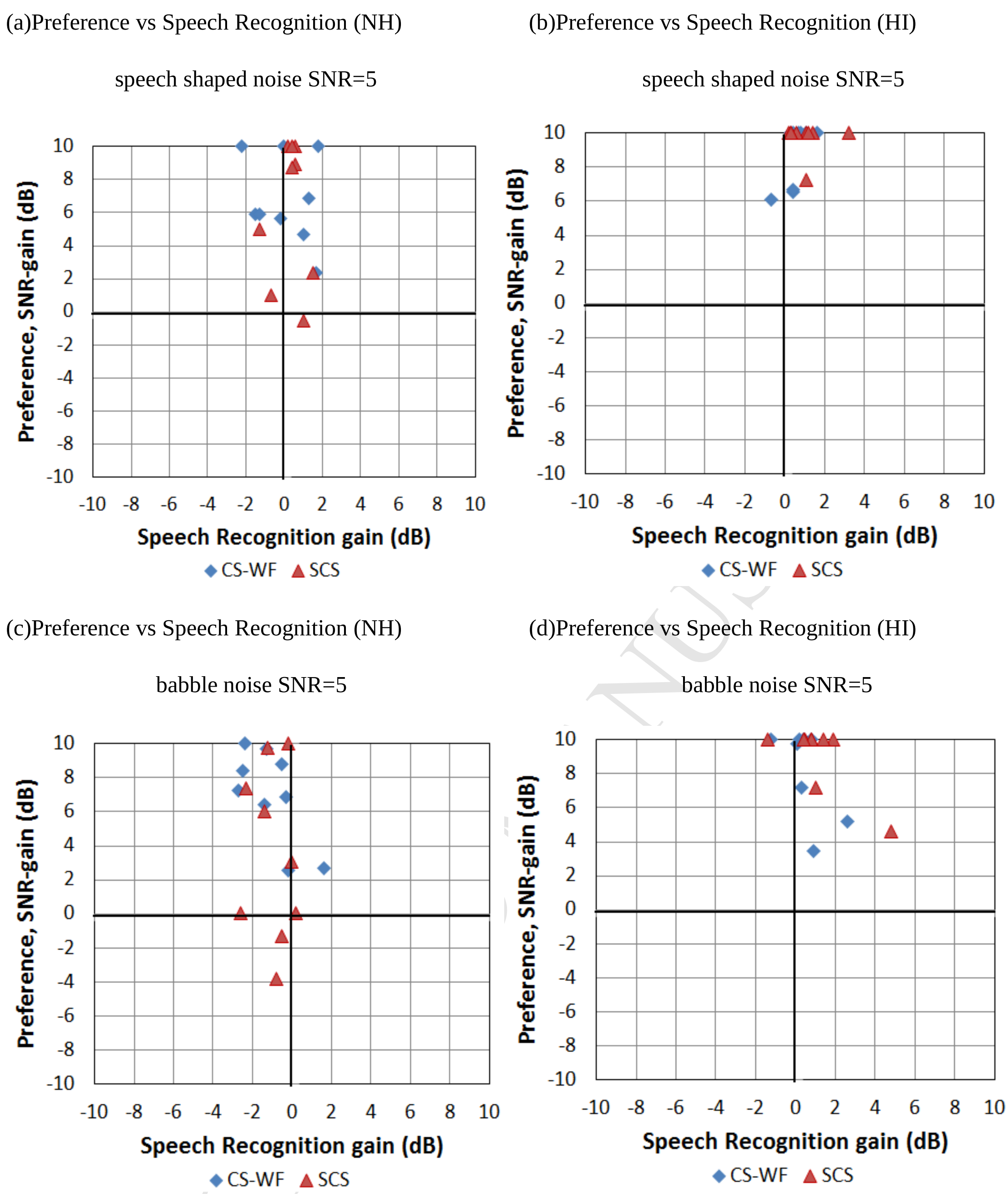

Fig. 8. Scatter plots showing individual 'preference' improvements vs. speech recognition gains in both noises (top vs. bottom) with both noise reduction algorithms (diamonds vs. triangles) with NH and HI listeners (left vs. right).

- A sparse coding shrinkage (SCS) algorithm on speech quality was evaluated.
- A method called Interpolated Paired Comparison Rating (IPCR) was adopted.
- The subjective measures were quantitatively compared with the objective measures.
- There was no large difference in quality between the SCS and the Wiener filtering.
- There was a difference in quality between hearing impaired and normal hearing.