EVALUATION OF A SPARSE CODING SHRINKAGE ALGORITHM IN NORMAL HEARING AND HEARING IMPAIRED LISTENERS

Jinqiu Sang[1], Hongmei Hu[1], Chengshi Zheng[2], Guoping Li[1], Mark E Lutman[1], Stefan Bleeck[1]

1. Institute of Sound and Vibration Research, University of Southampton, SO17 1BJ, Southampton, UK
2. Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, China

email: [js1e09, hongmei.hu, lgp, bleeck]@soton.ac.uk, mel@isvr.soton.ac.uk, cszheng@mail.ioa.ac.cn

ABSTRACT

Hearing impaired (HI) people struggle more than normal hearing (NH) listeners to understand speech in noisy environment. Previous evaluations of noise reduction algorithms on HI listeners have mainly concentrated on few algorithms like spectral subtraction or Wiener filtering. In this paper, a sparse coding shrinkage (SCS) noise reduction algorithm is proposed to compensate for some of the auditory deficits. The noise reduction performance by the SCS algorithm is compared with a Wiener filtering (CS-WF) approach, where the a *priori* signal-to-noise-ratio is estimated by the cepstral smoothing method. Speech recognition tests were performed to assess subjective intelligibility of SCS, CS-WF and noisy speech in babble noise and speech-shaped noise. Results show that both noise reduction algorithms have more potential to improve speech intelligibility in HI listeners than NH listeners; SCS provides more benefits than CS-WF for HI listeners especially in speech shaped noise.

Index Terms — sparse coding shrinkage, speech intelligibility, hearing impaired

1. INTRODUCTION

One of the most frequent complaints by hearing aid users is that hearing aids do not help to understand language in noisy environments. Hearing impaired (HI) listeners typically require a speech-to-noise ratio that is 3-6 dB higher than that of normal-hearing (NH) people to achieve the same level of speech intelligibility [1, 2]. To give HI listeners the same ability will require the development of more reliable and more efficient noise reduction algorithms in hearing aids.

Previous evaluations of noise reduction strategies in hearing aids mainly focused on spectral subtraction [2-5] or Wiener filtering algorithms [3]. However, most noise reduction algorithms were originally developed to improve speech perception for normal hearing subjects and were later adopted for hearing aid users. Because of hearing loss factors discussed below, algorithms that are optimal for NH listeners might not be optimal for HI listeners. Hearing loss factors include threshold elevation, loudness recruitment, reduced frequency selectivity and reduced temporal resolution [6, 7]. Automatic gain control (AGC) [8] compensates for threshold elevation and loudness recruitment, but there are currently no appropriate solutions to compensate for reduced frequency selectivity and reduced temporal resolution, which induces severe disruption of speech perception in noise.

A possible solution to the problem is to preserve less but key speech information while reducing the overall noise. This way, essential speech information is still present after the noise reduction, even factoring the reduced frequency selectivity and temporal resolution in HA users.

Here we propose a noise reduction strategy based on the principle of sparse coding shrinkage (SCS). It assumes a super-Gaussian (sparse) distribution [9] of the principal components in clean speech and SCS is performed in the principal components. SCS was first developed for image denoising [10] and has later been applied for speech enhancement [11-17]. Sparse coding has shown significant improvement in cochlear implant users [15] and this implies potential benefits of SCS in hearing aid users. It is also assumed that the principal components of the speech are super-Gaussian but difficult to calculate [18]. The shrinkage function in our SCS is therefore simplified by an approximation method.

Results of the noise reduction were evaluated by an adaptive speech recognition test with NH and HI listeners. Furthermore, SCS is compared with a Wiener filtering algorithm [19-21], as well as noisy speech in babble noise and speech shaped noise.

2. SPARSE CODING SHRINKAGE IN SPEECH

2.1. Implementation

Figure 1 shows the principle of calculating the sparse coding shrinkage in noisy speech. The noisy speech is transformed into principal components where clean signals are sparsely distributed and noise has Gaussian distribution.

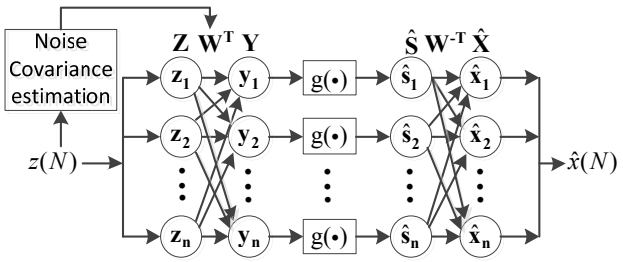

Figure 1 Flowchart of sparse coding shrinkage in noisy speech.

The shrinkage function g(·) is applied to suppress noisy components. After that an inverse transformation reconstructs the estimated clean speech signals.

The noisy speech signal z is assumed to be produced by corrupting the original speech sequence x with Gaussian noise:

$$z = x + n \tag{1}$$

The noisy speech matrix is constructed by reshaping z as overlapping frames (50% overlap)

$$\mathbf{Z} = \begin{bmatrix} z_1 & z_{m/2+1} & \cdots & z_{ml/2+1} \\ z_2 & z_{m/2+2} & \cdots & z_{ml/2+2} \\ \vdots & \vdots & \cdots & \vdots \\ z_m & z_{m/2+m} & \cdots & z_{ml/2+m} \end{bmatrix} \tag{2}$$

where l denotes the number of frames and m (m=64) is the 4-ms window in each column at a sampling rate of 16 kHz. After reshaping, the original noisy speech can be written as

$$\mathbf{Z} = \mathbf{X} + \mathbf{N} \tag{3}$$

The estimated clean speech covariance matrix and the estimated noise covariance matrix are denoted as $\hat{\mathbf{R}}_\mathbf{x}$ and $\hat{\mathbf{R}}_\mathbf{n}$ respectively (estimation details will be described in Section 2.3).

The transformation from noisy speech to principal components is realized by simultaneous diagonalization of the clean speech and noise covariance matrices [22] so that

$$\begin{aligned} \mathbf{W}^\mathbf{T}\hat{\mathbf{R}}_\mathbf{x}\mathbf{W} &= \mathbf{\Lambda}_\mathbf{x} \\ \mathbf{W}^\mathbf{T}\hat{\mathbf{R}}_\mathbf{n}\mathbf{W} &= \mathbf{I} \end{aligned} \tag{4}$$

Where $\mathbf{\Lambda}_\mathbf{x}$ and $\mathbf{W}$ are the eigenvalue matrix and eigenvector matrix respectively of the following matrix:

$$\mathbf{\Sigma} = \hat{\mathbf{R}}_\mathbf{n}^{-1}\hat{\mathbf{R}}_\mathbf{x} \tag{5}$$

When n is colored noise, pre-whitening is realized together with extraction of eigenvector matrix in Equation (5).

$$\mathbf{Y} = \mathbf{W}^\mathbf{T}\mathbf{Z} = \mathbf{W}^\mathbf{T}\mathbf{X} + \mathbf{W}^\mathbf{T}\mathbf{N} = \mathbf{S} + \mathbf{V} \tag{6}$$

where $\mathbf{Y} = [\mathbf{y_1}; \mathbf{y_2}; \cdots \mathbf{y_n}]$, $\mathbf{S} = [\mathbf{s_1}; \mathbf{s_2}; \cdots \mathbf{s_n}]$, $\mathbf{V} = [\mathbf{v_1}; \mathbf{v_2}; \cdots \mathbf{v_n}]$

The clean speech components $\mathbf{s_i}$ are in super-Gaussian distribution and noise components $\mathbf{v_i}$ are in Gaussian distribution. Therefore the sparse coding shrinkage function can be applied to each component $\mathbf{y_i}$ to estimate the clean components $\mathbf{s_i}$:

$$\hat{\mathbf{s}}_\mathbf{i} = \mathrm{g}(\mathbf{y_i}) \tag{7}$$

$\hat{\mathbf{S}} = [\hat{\mathbf{s}}_\mathbf{1}; \hat{\mathbf{s}}_\mathbf{2}; \cdots \hat{\mathbf{s}}_\mathbf{n}]$ is the estimated clean speech matrix in the space of principal components. Inverse transformation of the estimated clean matrix yields

$$\hat{\mathbf{X}} = \mathbf{W}^{-\mathbf{T}}\hat{\mathbf{S}} \tag{8}$$

Finally, the enhanced speech $\hat{x}$ is reconstructed by reshaping $\hat{\mathbf{X}}$ back into vector form by overlap and add method [23].

2.2 Super-Gaussian distribution and shrinkage function

The distribution of principal components of speech is assumed to be a linear combination of Gaussian and Laplacian distributions:

$$f_s(s_i) = C\exp\left(-as_i^2/2 - b|s_i|\right) \tag{9}$$

where C is an irrelevant scaling constant, different values of a and b represent different degrees of super-Gaussianity.

Through maximum-a-posterior (MAP) derivation [9], the shrinkage function corresponding to the distribution of (9) is derived as:

$$\hat{s}_i = \mathrm{g}(y_i) = \frac{1}{1+\sigma^2 a}\mathrm{sign}(y_i)\max\left(0, |y_i| - b\sigma^2\right) \tag{10}$$

$$b = \frac{2f_s(0)E\{s_i^2\} - E\{|s_i|\}}{E\{s_i^2\} - \left[E\{|s_i|\}\right]^2} \quad a = \frac{1}{E\{s_i^2\}}\left[1 - E\{|s_i|\}b\right]$$

where $f_s(0)$ is the value of the density function of s_i at zero, and σ^2 is the noise variance in the principal components.

This shrinkage function is interpolated between the shrinkage function of the Gaussian density and the shrinkage function of the Laplacian density. Specifically, when the distribution of s_i is Laplacian, a is 0 and b is estimated as $\sqrt{2/E\{s_i^2\}}$; when the distribution of s_i is Gaussian, b is 0 and a is estimated as $1/E\{s_i^2\}$. Therefore it is reasonable to constrain the values of a and b in the intervals of $\left[0, 1/E\{s_i^2\}\right]$ and $\left[0, \sqrt{2/E\{s_i^2\}}\right]$.

To simplify the estimation of a and b in Equation (10), we assume that

$a = \mu_1/E\{s_i^2\}$, $b = \mu_2 \times \sqrt{2/E\{s_i^2\}}$.

Here $\mu_{1,2}$ are coefficients to be adjusted experimentally. In our test, μ_1 is set to 1, μ_2 is set to 0.3, $E\{s_i^2\} = E\{y_i^2\} - \sigma^2$.

The choice of moderately super Gaussian distributions is justified by the criterion in [9] that when $\sqrt{E\{s_i^2\}} f_s(0) < 1/\sqrt{2}$, the distribution model can be assumed to be described as equation (9).

2.3. Noise estimation and pre-whitening

When the background noise is colored, the noise covariance matrix needs to be estimated as $\hat{\mathbf{R}}_n$ in Equation (5). Here, a noise estimation method proposed in [21] 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 noise covariance matrix is estimated by inverse Fourier transform of noise power spectral density according to the Wiener-Khinchin Formula [23].

$$\Phi_{nn}(e^{jw}) = \sum_{m=-\infty}^{+\infty} \phi_{nn}[m] e^{-jmw}$$
$$\phi_{nn}[m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_{nn}(e^{jw}) e^{jwm} dw \quad (11)$$

Where $\Phi_{nn}(e^{jw})$ is noise power spectral density and $\phi_{nn}[m]$ is the noise autocorrelation coefficients which can be derived through inverse Fourier transform of NPSD. When the mean of the noise is zero, noise auto-covariance coefficients equal noise auto-correlation coefficients. If the noise covariance matrix has the length of M, it can be constructed as a symmetric Toeplitz matrix with the first M values of the noise auto-covariance coefficients.

2.4. Comparison algorithm

This SCS algorithm is compared with a Wiener filtering approach, of which the code is provided by Timo Gerkmann [19-21]. This algorithm was chosen, because Wiener filters are used frequently in today's hearing aids and CS-WF is a competitive state of the art algorithm [19-21]. Since the a *priori* SNR of the WF approach is estimated by the cepstral smoothing method, we refer this approach as CS-WF herein. There are two critical techniques in CS-WF. One technique is to estimate 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 [21]. SCS also adopted the same NPSD estimation method as described in section 2.3. The other technique is that a priori SNR is estimated by temporal cepstrum smoothing with bias compensation [19, 20] where this technique could reduce the musical noise and suppress the non-stationary noise effectively.

3. EVALUATION

In order to evaluate the perceptual effects of the proposed SCS and the comparison CS-WF algorithms, we performed recognition tests through NH and HI listeners. Bamford-Kowal-Bench (BKB) [24] sentences recorded by a female British speaker were used as speech material. The sentence database comprised of 21 lists with 16 sentences in each list and 3 or 4 keywords in each sentence. Speech shaped noise (SSN) and babble noise were added in various quantities.

3.1. Subjective speech intelligibility tests

3.1.1 Participants

Eight NH listeners and three HI listeners with sensorineural hearing loss participated in this experiment. All subjects were native English speakers. The NH listeners had hearing thresholds at or below 20 dB HL from 250 Hz to 8 kHz (confirmed by PTA), and their ages ranged from 20 to 36. The 3 HI listeners all had moderate-to-severe sloping high frequency hearing losses. Tests were monaurally on the better ear. The audiograms of the tested ear of the HI listeners are shown in Figure 2. All the HI listeners were experienced hearing aid users and their ages ranged from 18 to 25. The tests were performed with their hearing aids taken off, and NAL procedure [8] was applied to each HI subject individually to compensate for hearing threshold elevation.

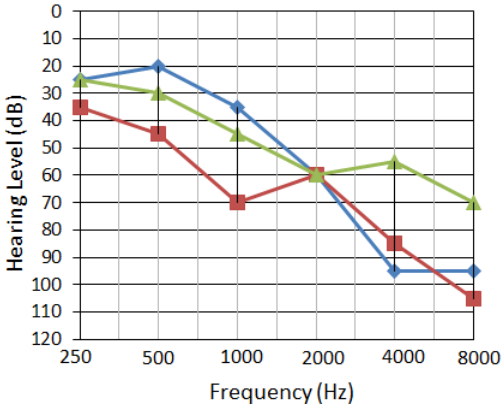

Figure 2 Audiograms of the tested ear of the 3 HI listeners.

3.1.2 Procedure

A total of six conditions were tested: two noise types (SSN, babble) and three noise reduction conditions ('noisy', SCS, WF). 'Noisy' indicates addition of noise without noise reduction strategies to show baseline performance. Sentences were randomly selected from the corpus for each test condition. Subjects were instructed to repeat as many words as they could after listening to each sentence with no feedback given during the tests. For familiarization, participants practiced the procedure with one randomly selected condition. The order of the six conditions was

randomized but balanced among the listeners using a latin square. Each experiment took around half an hour.

A three-up one-down adaptive procedure was used to find the speech-to-noise ratio required for 79.4% correct recognition in each condition [5], which is defined as speech reception threshold (SRT) in dB. A sentence was deemed to have been recognised correctly when at least two keywords were repeated correctly. Sentence order was controlled so that participants did not receive the same sentence twice. The step size was 1 dB. All sound files from -10 dB to 15 dB, with and without noise reduction algorithms in different SNRs, were pre-processed offline.

All listeners were seated in a sound-isolated room and listened to the sounds presented through a Sennheiser HDA 200 headphone 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 comfortable listening level.

3.1.3 Subjective experimental results

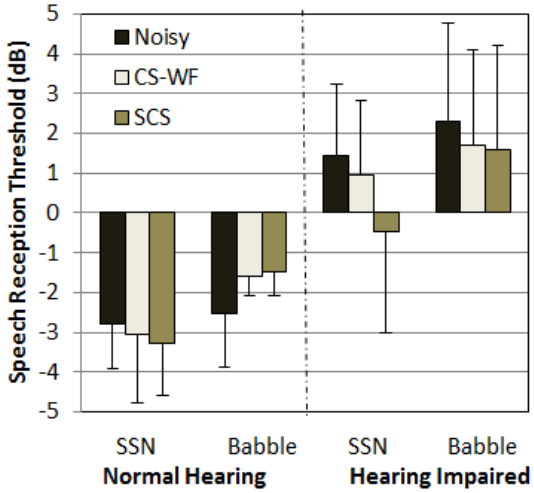

Figure 3 SRTs for different conditions in normal hearing and hearing impaired listeners. SSN: speech shaped noise; Noisy: noisy speech without noise reduction algorithms; WF: Wiener filtering; SCS: sparse coding shrinkage. Error bars show 1 standard deviation.

Figure 3 shows the average performance of all participants in all six test conditions: SSN-Noisy, SSN-WF, SSN-SCS, Babble-Noisy, Babble-WF, Babble-SCS, for both NH (left) and HI subjects (right). For NH subjects, the effect of noise reduction algorithm is not significant [$F(2,14)=1.33$, $p>0.05$], but the effect of noise type is significant for NH subjects [$F(1,7)=8.13$, $p<0.05$]. There is no interaction between noise type and noise reduction algorithm [$F(2,14)=2.74$, $p>0.05$]. These results are in accordance with previous evaluations [25] showing that most single channel noise reduction algorithms cannot improve speech intelligibility in NH listeners. For HI subjects, neither the noise reduction algorithm nor the noise type has a significant effect [$F(2,4)=7.82$, $p>0.05$, and $F(1,2)=3.28$, $p>0.05$, respectively]. There is no interaction between noise reduction algorithm and noise type [$F(2,4)=5.74$, $p>0.05$]. Because of the limited number of HI subjects in the experiments so far, the investigation is underpowered. At this state, we cannot say for sure if there is no effect of processing algorithm or noise type. Further experiments involving more subjects will clarify this point. Individual inspection of results shows that all three HI subjects had better speech intelligibility with both noise reduction algorithms. This implies that HI listeners might benefit more from noise reduction strategies than NH listeners. In general, this might be explained by the fact that noise reduction algorithms distort the speech to some degree. Due to suprathreshold deficits like reduced frequency selectivity, HI subjects are more used to listen to distorted speech than NH subjects. This is also in accordance with [26] where it was shown that speech intelligibility degrades significantly in NH subjects when speech is distorted compared to HI subjects. This is also in accordance to the finding that noise reduction schemes based on the ideal binary mask could benefit HI listeners more than NH listeners [27].

For HI listeners, SCS shows slightly higher intelligibility improvement, especially in speech shaped noise than CS-WF. This motivates us to further develop noise reduction strategies that are optimized for HI listeners. Sparse stimuli could compensate further for suprathreshold deficits. The advantage of noise reduction using sparse stimuli have already been demonstrated for listeners who are profoundly hearing impaired and use a cochlea implant [14-16].

The error bars in Fig. 3 illustrate that the hearing impaired subjects show correspondingly large inter-subject variability compared to normal hearing subjects. We assume that this is due to individually different auditory deficits and individual experience with hearing aids. It is expected that if we were able to give hearing impaired participants more practice with new noise reduction algorithms, they would benefit even more from the sparse noise reduction algorithm in terms of speech intelligibility.

4. CONCLUSIONS

Although the number of participants in our experiments was small, we conclude that noise reduction strategies hold more promise to help HI subjects than NH subjects in severe noise environments. Further experiment with higher number of participants will clarify this point. In informal questions, HI subjects reported that both noise reduction algorithms improve speech quality.

Compared to the CS-WF algorithm, SCS has the potential to bring more benefits to hearing impaired subjects especially in stationary noise, such as speech shaped noise.

Although the difference is not significant in listening experiments, the trend is visible.

We conclude that noise reduction algorithms that consider auditory deficits probably help HI listeners more than noise reduction algorithms that were originally developed for NH listeners. With more practice, performance probably further improves when HI listeners learn to adapt to the increased speech distortion in SCS. This motivates future research in noise reduction algorithms that take account of auditory deficits rather than simply adopting noise reduction algorithms from common telecommunication systems.

5. ACKNOWLEDGMENT

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 subjects. This work was supported by the European Commission within the ITN AUDIS (grant agreement number PITN-GA-2008-214699).

6. REFERENCES

[1] R. Plomp, "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," Ear Hear, vol. 15, pp. 2-12, 1994.

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

[3] H. Levitt, M. Bakke, J. Kates, A. Neuman, T. Schwander, and M. Weiss, "Signal processing for hearing impairment," Scandinavian Audiology, Supplementum, vol. 38, pp. 7-19, 1993.

[4] C. Elberling, C. Ludvigsen, and G. Keidser, "The design and testing of a noise reduction algorithm based on spectral subtraction," Scandinavian Audiology, Supplementum, vol. 1993, pp. 39-49, 1993.

[5] M. Dahlquist, M. E. Lutman, S. Wood, and A. Leijon, "Methodology for quantifying perceptual effects from noise suppression systems," Int. J. Audiol., vol. 44, pp. 721-732, 2005.

[6] Y. Nejime and B. C. J. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," The Journal of the Acoustical Society of America, vol. 102, pp. 603-615, 1997.

[7] H. Hu, J. Sang, and M. E. Lutman, "Simulation of hearing loss using compressive gammachirp auditory filters," in ICASSP, Prague, Czech Republic, 2011.

[8] H. Dillon, Hearing Aids. Thieme, New York, 2001.

[9] A. Hyvarinen, "Sparse code shrinkage: Denoising of nonGaussian data by maximum likelihood estimation," Neural Computation, vol. 11, pp. 1739-1768, 1999.

[10] A. Hyvarinen, P. Hoyer, and E. Oja, "Sparse code shrinkage for image denoising," in Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence., 1998, pp. 859-864 vol.2.

[11] X. Zou, P. Jancovic, L. Ju, and M. Kokuer, "Speech Signal Enhancement Based on MAP Algorithm in the ICA Space," IEEE Trans. Signal Processing, vol. 56, pp. 1812-1820, 2008.

[12] I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Speech enhancement using the sparse code shrinkage technique," in ICASSP, 2001, pp. 621-624 vol.1.

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

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

[15] G. Li and M. E. Lutman, "Sparse Stimuli For Cochlear Implants," in Proc. EUSIPCO, Lausanne, Switzerland, 2008.

[16] H. Hu, G. Li, L. Chen, J. Sang, S. Wang, M. E. Lutman, and S. Bleeck, "Enhanced sparse speech processing strategy for cochlear implants," in Eurpsipco, Barcelona, Spain, 2011.

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

[18] J. Jesper and H. Richard, "Improved Subspace-Based Single-Channel Speech Enhancement Using Generalized Super-Gaussian Priors," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, pp. 862-872, 2007.

[19] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp. 4897-4900.

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

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

[22] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted with colored noise," IEEE Transactions on Speech and Audio Processing, vol. 11, pp. 334-341, 2003.

[23] J. R. Deller, J. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals. New York: IEEE Press, 2000.

[24] J. Bench, A. Kowal, and J. Bamford, "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," British Journal of Audiology, vol. 13, pp. 102-112, 1979.

[25] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am., vol. 122, pp. 1777-1786, 2007.

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

[27] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am., vol. 125, pp. 2336-2347, 2009.