IMPROVED FACE-TO-FACE COMMUNICATION USING NOISE REDUCTION AND SPEECH INTELLIGIBILITY ENHANCEMENT

Anthony Griffin[1,2]*, Tudor-Cătălin Zorilă*[2]*, Yannis Stylianou*[2,3]

[1]AUT University, Auckland, New Zealand
[2]University of Crete, Heraklion, Greece
[3]Toshiba Cambridge Research Laboratory, United Kingdom

agriffin@aut.ac.nz, ztudorc@gmail.com, yannis@csd.uoc.gr
yannis.stylianou@crl.toshiba.co.uk

ABSTRACT

Significant improvements in intelligibility of speech in noise can be obtained by modifying the speech signal in the time and/or frequency domains. However, most speech intelligibility enhancement algorithms are designed to use clean speech as an input, and their performance suffers once the input speech signal-to-noise ratio decreases, a common case in face-to-face communication environments such as restaurants or cafés. In this work we investigate whether a particularly successful speech intelligibility enhancement system—spectral shaping and dynamic range compression—and various front-end noise reduction methods might be suitable in such environments. Our evaluations suggest that such a complete system would provide an increase in speech intelligibility equivalent to a gain of 10 dB input signal-to-noise ratio in the more challenging face-to-face communication environments.

Index Terms— speech intelligibility, speech enhancement, noise reduction, face-to-face communication.

1. INTRODUCTION

Hearing and understanding speech is an extremely important part of a person's ability to communicate with others. As the background noise increases it gets harder and harder to make out the content of the speech of interest. This is true for a person with normal hearing, but even more so for a hearing impaired person.

In order to combat this, many signal processing algorithms have been developed to increase the intelligibility of speech in noise (see [1] and the references within). One of these algorithms—spectral shaping and dynamic range compression (SSDRC) [2]—has been shown to be a state-of-the-art method to improve speech intelligibility [1].

However, most speech intelligibility enhancement algorithms are designed to use clean speech as an input, and their performance suffers once the input speech signal-to-noise ratio decreases, a common case in face-to-face communication environments such as restaurants or cafés. SSDRC is not immune to this effect, and suffers similarly when the background noise increases.

A potential solution to this may be to pre-process the noisy speech signal to reduce the noise before passing it to SSDRC. Speech enhancement in the presence of noise is an area that has been investigated for many decades [3–5]. The great majority of these methods analyze the speech signal in the frequency domain, and process the amplitude spectrum, ignoring the phase. Much more recently, methods have been developed to include the phase in this processing, providing very encouraging results [6–10].

In this paper, we develop a face-to-face communication system that uses the phase-aware methods of [6–10] as a pre-processing input stage to SSDRC, increasing the signal-to-noise ratio (SNR) of the speech passed to SSDRC. Our system can then accept noisy speech as an input, and output a modified speech signal that has greatly enhanced intelligibility.

2. SSDRC

Spectral shaping and dynamic range compression (SSDRC) [2] has been shown to be an excellent method to improve speech intelligibility [1]. The spectral shaping operates in the frequency domain, and the dynamic range compression (DRC) operates primarily in the time domain. The spectral shaping consists of two cascaded subsystems which are adaptive to the probability of voicing: (i) an adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. Furthermore, a third fixed spectral shaping is used to prevent attenuation of high frequencies in the speech signal during signal reproduction.

The output of the spectral shaping system is then input to the DRC, which has a dynamic and a static stage. During the dynamic stage, the envelope of the total time signal is dy-

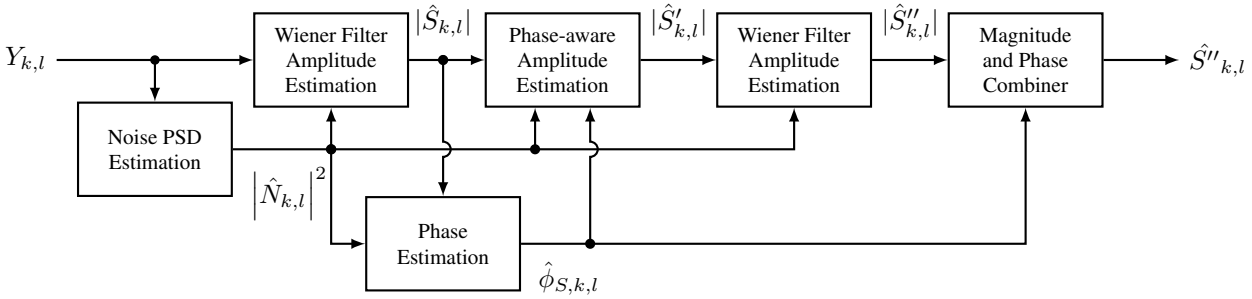


Fig. 1. The noise reduction subsystem.

namically compressed with a 2 ms release time constant and almost instantaneous attack time constant. During the static amplitude compression, an input-output envelope characteristic (IOEC) is applied to the dynamically compressed envelope converted to dB. The 0 dB reference level is set to 0.3 times the peak of the signal envelope. Thus, DRC enhances the transient components of speech.

There is a final stage in SSDRC that ensures the input power and the output power are the same, this guarantees that the gains in intelligibility are not due to signal amplification.

3. PROPOSED SYSTEM

As SSDRC was designed to operate on clean speech, any noise on the input speech signal will be treated as if it was part of the speech, leading to unwanted masking of the speech, and reduced intelligibility. In order to combat this loss in intelligibility, we propose the use of a preprocessing input stage to reduce the noise, effectively increasing input speech's SNR. Although this type of processing is often referred to in the literature as *speech enhancement*, here we will refer to it as *noise reduction* to differentiate it from the speech enhancement of the following SSDRC stage.

3.1. Noise Reduction

Most noise reduction methods operate only on the amplitude spectrum of the noisy speech, and ignore the phase. Perhaps the most classic example of this is the Wiener filter [3], which provides a minimum mean square error (MMSE) estimate of the amplitude spectrum of the speech signal.

After the appropriate sampling and conversion to the Fourier domain, we can write the contents of the k-th bin of the l-th frame of the noisy speech signal as

$$Y_{k,l} = S_{k,l} + N_{k,l}, \tag{1}$$

where $S_{k,l}$ and $N_{k,l}$ are the speech and noise in the k-th bin of the l-th frame, respectively. The softmask gain of the Wiener filter can then be given as

$$G_{k,l} = \frac{\left|\hat{S}_{k,l}\right|^2}{\left|\hat{S}_{k,l}\right|^2 + \left|\hat{N}_{k,l}\right|^2}, \tag{2}$$

where the estimate of the magnitude of the speech signal $\left|\hat{S}_{k,l}\right|$ is often found by recursion, and the estimate of the magnitude of the noise power spectral density (PSD) $\left|\hat{N}_{k,l}\right|^2$ may be found by methods such as [11] or [12]. Once $G_{k,l}$ has been calculated, the speech estimate is given by

$$\hat{S}_{k,l} = G_{k,l}\left|Y_{k,l}\right| \exp\left(\mathrm{i}\phi_{Y_{k,l}}\right), \tag{3}$$

where $\phi_{Y_{k,l}}$ is the phase of the Fourier domain noisy speech signal, given by

$$\phi_{Y_{k,l}} = \angle Y_{k,l}. \tag{4}$$

An obvious next step is to consider accurately estimating the phase of the speech signal, but until recently this had proved elusive. However, recently work in [6] proposed an effective method to estimate this phase $\hat{\phi}_{S_{k,l}}$, based on geometry and group delay minimization. An improved estimate of the speech signal can then be formed as

$$\hat{S}'_{k,l} = G_{k,l}\left|Y_{k,l}\right| \exp\left(\mathrm{i}\hat{\phi}_{S_{k,l}}\right). \tag{5}$$

Other work in [8] and [10] then uses the initial estimate of the magnitude of the speech signal $\left|\hat{S}_{k,l}\right|$ and the estimate of the phase of the speech signal $\hat{\phi}_{S_{k,l}}$ to produce a $phase-aware$ estimate of the magnitude of the speech signal $\left|\hat{S}'_{k,l}\right|$. This can then be used in another Wiener filter to produce a phase-aware softmask gain as

$$G'_{k,l} = \frac{\left|\hat{S}'_{k,l}\right|^2}{\left|\hat{S}'_{k,l}\right|^2 + \left|\hat{N}_{k,l}\right|^2}. \tag{6}$$

Finally, similar to (3) and (5), the estimated speech signal is given by

$$\hat{S}''_{k,l} = G'_{k,l}\left|Y_{k,l}\right| \exp\left(\mathrm{i}\hat{\phi}_{S_{k,l}}\right). \tag{7}$$

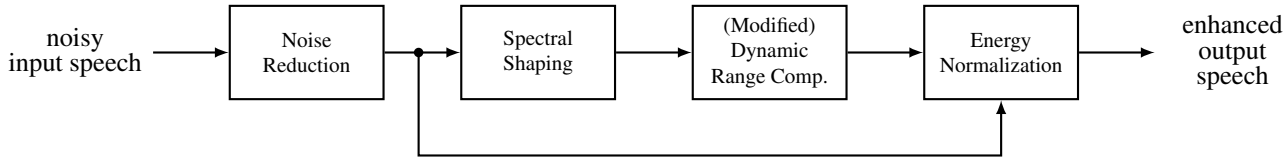


Fig. 2. The proposed system, noise-tolerant SSDRC.

The full noise reduction system is illustrated in Fig. 1. The noisy speech signal $Y_{k,l}$ is first fed into a Noise PSD Estimation block which produces an estimate of the noise PSD $\left|\hat{N}_{k,l}\right|^2$, which is used by most of the other blocks. The first Wiener filter then produces an initial estimate of the amplitude of the speech signal $\left|\hat{S}_{k,l}\right|$, and this is used to generate an estimate of the phase of the speech signal $\hat{\phi}_{S_{k,l}}$. The Phase-aware Amplitude Estimation block then uses $\hat{\phi}_{S_{k,l}}$ to produce an improved estimate of the amplitude of the speech signal $\left|\hat{S}'_{k,l}\right|$, which is further refined by the second Wiener filter to produce $\left|\hat{S}''_{k,l}\right|$. This is then combined with $\hat{\phi}_{S_{k,l}}$ to produce the final estimate of the speech signal.

3.2. Modification of SSDRC

After initial testing, it became apparent that the part of SSDRC that was the most sensitive to noise on the input speech was the DRC. This is because the DRC is intended to transfer energy over time from louder speech segments (such as voiced speech) to quieter (often unvoiced) parts of the speech signal, resulting in short passages of noise being amplified and consequently, reduced intelligibility. The solution we pursued was a modification of the IOEC curve [2]. Essentially this results in a change to the threshold of silence of the IOEC curve based on the input SNR γ. Let the threshold of silence be denoted ξ. In the clean speech scenario, $\gamma = \infty$, and $\xi = -30$ dB. We then vary ξ as:

$$\xi = \begin{cases} 0 \text{ dB}, & \text{if} \quad \gamma \leq 0 \text{ dB} \\ -\gamma, & \text{if} \quad 0 \text{ dB} < \gamma \leq 30 \text{ dB} \\ -30 \text{ dB}, & \text{if} \quad \gamma > 30 \text{ dB} \end{cases}. \tag{8}$$

3.3. Noise-tolerant SSDRC

The final proposed system which we refer to as noise-tolerant SSDRC (ntSSDRC), is shown in Fig. 2. The noisy input speech is fed into the noise reduction subsystem described in Section 3.1 and Fig. 1, output of this is then processed by the spectral shaper which enhances the frequency characteristics of the speech. This output is then modified in the time domain by the DRC stage with the input SNR-dependent modification of Section 3.2. Finally, the energy of the output signal is constrained to be the same as that at the input to the spectral shaper.

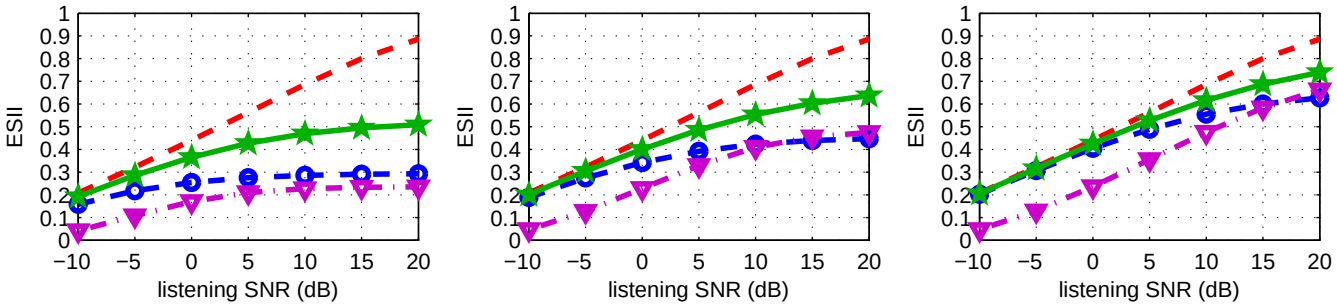
4. DISCUSSION

The final goal of this work was of course improved speech intelligibility—itself a very subjective measurement—and the proposed system of Figures 1 and 2 is obviously highly non-linear. Furthermore, speech intelligibility is somewhat both speech sample- and speaker-dependent. Thus, all the evaluations in this work were done on 5 different speakers (three males, two females) each saying 5 different sentences, for a total of 25 different samples. The clean speech was recorded at 16 kHz, to which the appropriately scaled speech-shaped noise was added. The extended speech intelligibility index (ESII) [13] was used as the final performance indicator. Nonetheless, the perceptual evaluation of speech quality (PESQ) [14] was used at intermediate stages to help tune and test the system, as well as informal intelligibility listening tests.

The overall effect of each subsystem in a system like this is hard to predict, so changes have to be carefully considered. Indeed, optimizing each block in the whole system independently will almost certainly *not* result in the best overall performance. So each block's changes had to be checked against its impact on the overall performance.

When it came to choosing parameters for the noise reduction subsystem, there was a delicate balance to be maintained between eliminating as much of the noise in the pauses between speech as possible, and minimizing speech distortion. For instance, for the noise PSD estimation we chose to use [12] over [11] as the latter had a tendency to underestimate the noise, which in turn reduced the amount of noise that was removed from the noisy speech signal. The majority of performance improvement in ESII can be obtained by using the output of the first Wiener filter to generate an estimate of the speech signal as in (3). The remaining blocks of the noise reduction system further improve the ESII, but at significant computational cost. In particular, the Phase Estimation is extremely computationally intensive.

We also investigated the possibility of smoothing the gain of the final Wiener filter in the cepstral domain following the work of [15]. Although this did decrease the so-called "musical noise" in the output of the noise reduction subsystem, it also affected the speech in such a way that the final output of the ntSSDRC system had a worse ESII, and we chose not to use it.



(a) input SNR of 0 dB (b) input SNR of 10 dB (c) input SNR of 20 dB

Fig. 4. Extended Speech Intelligibility Index for various input SNRs and types of processing. The dashed red lines are always SSDRC applied to clean speech. The other lines have the specified input SNR. The solid green lines with stars are ntSSDRC applied to noisy speech, the blue dashed lines with circles are SSDRC applied to noisy speech, and the dashed magenta lines with triangles are plain noisy speech.

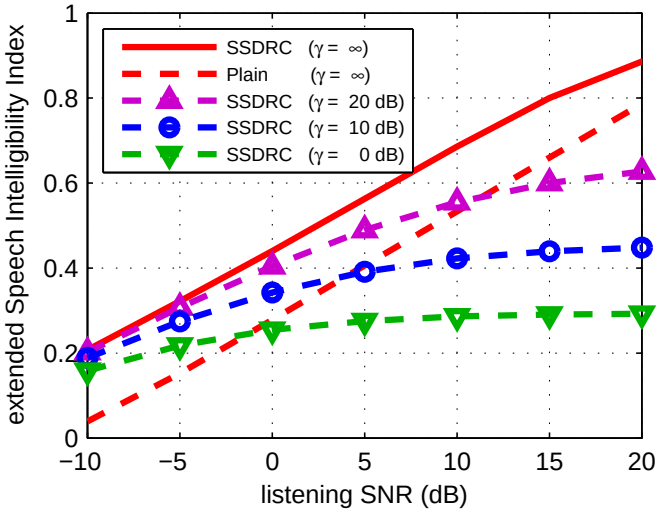


Fig. 3. Extended speech intelligibility index between plain and SSDRC speech for differing levels of input SNR (γ).

The modifications to DRC are promising but further investigation is required here to integrate them properly with the noise reduction. This due to the fact that the characteristics of the noise change significantly after the noise reduction block.

5. RESULTS

5.1. Performance

We first present the results that were the motivation for this work, a demonstration of how the ESII of SSDRC speech deteriorates as the input speech SNR γ decreases. This is clearly shown in Fig. 3. In particular, the losses at $\gamma = 0$ dB and $\gamma = 10$ dB are significant.

The performance of ntSSDRC is shown in Fig. 4, where it is clear that the use of the noise reduction subsystem of ntSSDRC allows it to regain some of the lost performance due to noisy input speech. The gains in intelligibility are greatest at the lower input SNR values, especially at 0 dB where it is most needed. Overall, it is evident that the use of ntSSDRC provides the same ESII as SSDRC with an input SNR 10 dB higher. Thus we claim that ntSSDRC should provide a 10 dB gain in intelligibility. Fig. 4 also highlights the gains of ntSSDRC over plain noisy speech.

5.2. Computational Complexity

A secondary goal of this work was the development of a working face-to-face communication system. In order to test its real-time performance, the proposed system was implemented in C++, on a Windows 8 laptop with a Core i7 processor running at 2.4 GHz. With input and output soundcard frame sizes of 424 samples at 44.1 kHz, less than 20% of the available processing time between audio interrupts was needed to perform the processing of the proposed system, confirming its suitability as a real-time system. Note that the majority (over half) of the processing time is taken by the phase estimation algorithm of the noise reduction subsystem.

6. CONCLUSIONS

We have presented the first work that considers combining state-of-the-art noise reduction algorithms with state-of-the-art speech intelligibility enhancement techniques to provide a face-to-face communication system designed to work in a significantly noisy environment. In particular, our system is suitable for real-time application, and our evaluations are very encouraging, suggesting that our system will provide a 10 dB gain in intelligibility over a baseline system without noise reduction. Our next step will be to perform extensive speech intelligibility listening tests.

7. REFERENCES

[1] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[2] T.C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[3] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[7] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[8] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[9] P. Mowlaee, M.K. Watanabe, and R. Saeidi, "Phase-Aware Single-Channel Speech Enhancement," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2013.

[10] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 2–5, 2013.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, 2001.

[12] T. Gerkmann and R.C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[13] K.S. Rhebergen and N.J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners.," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[14] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[15] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.