SPEECH-IN-NOISE INTELLIGIBILITY IMPROVEMENT BASED ON POWER RECOVERY AND DYNAMIC RANGE COMPRESSION

Tudor-Cătălin Zorilă[1], *Varvara Kandia*[2], *Yannis Stylianou*[2]

[1]Telecommunication Department, Politehnica University of Bucharest (UPB), Romania
[2]ICS-FORTH and Computer Science Department, University of Crete, Heraklion, Crete, Greece
ztudorc@gmail.com, vkandia@ics.forth.gr, yannis@csd.uoc.gr

ABSTRACT

The ability to detect speech in noise plays a significant role in our communication with others. In this work we suggest to modify the original speech signal before this is presented in the noisy environment by combining a signal to noise ratio recovery approach with dynamic range compression in order to improve the intelligibility of the speech in noise. The modification is performed under the constraint of equal global signal power before and after modifications. Experiments with speech shaped (SSN) and competed speaker (CS) types of noise at various low SNR values, show that the suggested approach outperforms state-of-the art methods in terms of the Speech Intelligibility Index (SII) as well as in informal listening tests. Comparing with a state-of-the-art method there is an improvement of 4 dB and 8 dB in terms of SNR gain, for the SSN and the CS types of noise, respectively.

Index Terms— speech in noise;speech modifications;SII;speech intelligibility

1. INTRODUCTION

Speech produced under real conditions (not in a recording studio, nor in a quiet room) is not always intelligible due to the presence of background noise. This noise may mask part of the speech signal such that not all speech information is available to the listener. The ability to detect speech in noise plays a significant role during a conversation, or understanding an announcement at the airports or train stations. Regarding conversations, it has been observed that there is an involuntary tendency of speakers to increase their vocal effort when speaking in loud noise to enhance the audibility of their voice (Lombard effect or Lombard reflex [1]). In telecommunications, it would be very beneficial for the listeners if the phone devices could automatically detect the noise environment of the listener and modify accordingly talker's speech in order to improve the intelligibility of the transmitted signal. This is referred to as *near end* listening enhancement problem [2]. Similar problems exist in broadcasting where pre-transmission enhancement techniques are applied on the baseband audio signal. In all these cases, the noise signal can not be modified, simply because it belongs to the environment where the listener is located. The remaining option, and assuming a speech reproduction system (i.e., announcements, text-to-speech synthesis) or transmission channel (i.e., telephone), is then to manipulate the produced speech signal in order to improve its intelligibility for the listener.

In a series of papers, B. Sauert and P. Vary suggested many speech enhancement approaches for improving the intelligibility of speech in noise conditions assuming the noise is known [2] [3]. In [2], the speech enhanced algorithm raised the average speech spectrum over the average noise spectrum in order to recover a target signal-to-noise-ratio. In [3], the enhancement algorithm is optimized with respect to the Speech Intelligibility Index (SII) [4], under the constraint of an unchanged average power of the speech signal. In very early studies for near-end speech enhancement, Niederjohn and Groteluesche suggested a rapid amplitude compression following high-pass filtering for processing speech before its reception by the listener [5].

For clear, and Lombard speech it has been reported that there is higher energy in the mid-frequency region of the frequency spectrum [6] [7] [8] [9] comparing with casual and non Lombard speech [1].

In the work of Hazan and Simpson, it has been shown that selective reinforcement of bursts and vocalic onsets and offsets can provide significant improvements to the intelligibility of the subsequently degraded speech signal, even for the same overall signal-to-noise ratio [10]. Enhancement of the transient components of speech has also been shown to improve intelligibility of speech in noise conditions [11].

In this work we consider improving the intelligibility of speech in noise, by combining previous research attempts and observations into one system, under the constraint of equal signal power before and after the modification. We revisit the earlier work in this domain ([5]) acknowledging that the high-pass filtering of speech indeed improves intelligibility of speech in white noise. However, instead of filtering the un-

[1]other differences include longer and more frequent pauses, reductions in speaking rate and expansions of the vowel space and modifications (especially for clear speech)

modified speech signal as suggested in [5], we apply the high-pass filter on the output signal of a system which tries to recover a good signal-to-noise ratio as this was suggested in [2]. This has the following advantages over previous works. First, it takes into account the noise characteristics which is not the case in [5], suggesting therefore an optimum frequency reallocation of the energy given the noise spectrum and a prespecified SNR ratio gain. Second, the high-pass filter removes the low frequencies which, in voiced speech, contain most of the energy of the signal. This doesn't decrease the intelligibility of the high-pass filtered speech, while makes easier to maintain the initial SNR without reducing significantly the optimum energy of the signal in the pass band frequencies [2]. This high-pass process is absent form the work of Sauert et al. [2].

Following the high-pass filtering, and based on the observations of Hazan et al. [10], we suggest the use of a dynamic range compression (DRC) algorithm which reduces the peak-to-rms ratio of the input signal. This has as effect to reduce the energy of the sonorant speech segments (i.e. mostly voiced) and increase the energy for voiced offsets and onsets, for bursts and for fricatives. This last step actually re-allocates energy over time. Experiments with speech shaped (SSN) and competed speaker (CS) types of noise at various low SNR values, show that the suggested overall system outperforms state-of-the art methods in terms of SII as well as in informal listening tests. Considering that constraints in the power of modified speech are imposed, we show that there is an improvement of 4 dB and 8 dB in terms of SNR gain, for the SSN and the CS types of noise, respectively.

The rest of the paper is organized as follows. In Section 2 we present the energy reallocation algorithm in the frequency domain where the SNR-recovery algorithm is combined with a high-pass filer. Section 3 describes the Dynamic Range Compression for the reallocation of the signal energy over time. Experiments with two types of noise, SSN and CS, are described in Section 4 and finally, Section 5 concludes the paper.

2. ENERGY REALLOCATION IN FREQUENCY

In this section we will shortly review the SNR-recovery algorithm suggested in [2] since it is the first part of the suggested system.

The SNR-recovery algorithm suggests to enhance the magnitude spectrum, $S(m,\Omega_k)$ of frame m (we assume a frame-by-frame processing) as following:

$$\hat{S}(m,\Omega_k) = G(m,\Omega_k)S(m,\Omega_k) \tag{1}$$

where

$$G(m,\Omega_k) = min\left\{max\left\{\sqrt{\xi\frac{\Phi_{NN}(m,\Omega_k)}{\Phi_{SS}(m,\Omega_k)}},1\right\},G_{max}\right\} \tag{2}$$

denotes the gain per frequency Ω_k, Φ_{NN} and Φ_{SS} denote the short-term power spectra density (PSD) of the noise and the speech signal respectively, while ξ and G_{max} denote the desired target SNR and the maximum allowed gain, respectively. The short-term PSD for the speech and the noise is computed as the recursive average of their periodograms, $|S(m,\Omega_k)|^2$, $|N(m,\Omega_k)|^2$. As an example, for speech:

$$\Phi_{SS}(m,\Omega_k) = a_S\Phi_{SS}(m-1,\Omega_k) + (1-a_S)|S(m,\Omega_k)|^2 \tag{3}$$

where $a_S \in [0,1]$. Same recursive formula is used for the estimation of $\Phi_{SS}(m,\Omega_k)$ but with a different time constant a_N. Finally, the maximum value of the modified spectrum, $\hat{S}(m,\Omega_k)$ is limited by a third gain $\hat{S}_{max}(\Omega_k)$. In this work we followed the recommendations for these variables provided in [2]. Therefore, $\xi = 15dB$, $G_{max} = 30dB$, $\hat{S}_{max}(\Omega_k) = 120dB$, $\forall k$, the time constants were set to: $a_S = 0.996$ and $a_N = 0.96$.

Following the work of Niederjohn et al. [5], a high pass filter is applied in the frequency domain:

$$\breve{S}(m,\Omega_k) = \begin{cases} 0, & if\ \Omega_k \le \Omega_c \\ \hat{S}(m,\Omega_k), & if\ \Omega_k > \Omega_c \end{cases} \tag{4}$$

By listening tests, it was found that if $\Omega_c = 800Hz$, the high-pass filtered speech signal does not loose much of its intelligibility while there is a lot of energy savings because of the elimination of the low frequency components.

Fig. 1 shows an example of magnitude spectra for the original speech and the noise (SNR is -4 dB), as well as the magnitude spectrum of the original speech after applying the SNR-recovery algorithm and its high-passed version. It is worth noticing that keeping only the frequencies beyond 800 Hz, there is a saving of 98% in energy (i.e, the new magnitude spectrum contains only the 2% of the energy of the SNR-recovered signal). In Fig. 2, the original signal (upper panel), and the reconstructed high-passed SNR-recovered signal (lower panel) are depicted. Since the SNR-recovered algorithm does not put any constraint in the total energy of the modified signal, the reconstructed signal has considerably more energy than the original one. Analysis and synthesis of signals is performed in frame-by-frame basis, and the output signal is reconstructed by Overlap and add (OLA).

3. ENERGY REALLOCATION IN TIME

For the reallocation of the energy over time, the use of a Dynamic Range Compression (DRC) system is suggested. The output from the SNR-recovery and the high-pass filter is the input to DRC.

The goal of DRC is to produce a time-varying gain to reduce the envelope of the signal variations. This gain is derived from a desired input/output envelope characteristic (IOEC). The IOEC used in this work is shown in Fig. 3 with its three characteristic zones; unity gain, expansion, and compression.

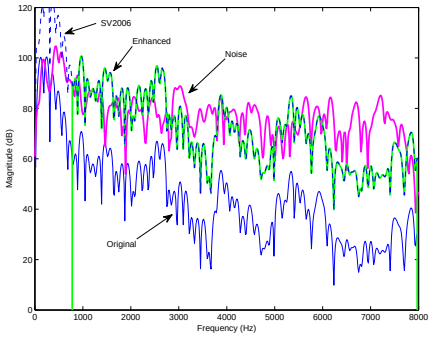

Fig. 1. Example of magnitude spectra for the original speech (blue solid), noise (magenta solid), enhanced by the SNR-recovery algorithm [2] (blue dashed, labeled as SV2006), and its high-passed version (green solid). Initial SNR: - 4.5 dB.

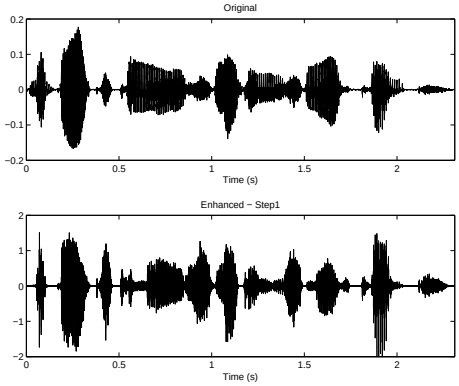

Fig. 2. Original speech (upper panel) and high-passed SNR-recovery signal (lower panel)

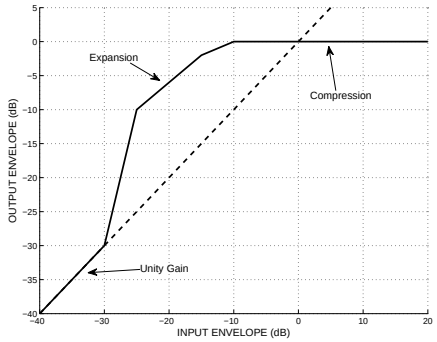

Fig. 3. Input-Output Envelope Characteristic (IOEC) curve.

The envelope of the speech signal, $s(n)$ is computed as the magnitude of the analytic signal:

$$r(n) = s(n) + j\check{s}(n)$$

where $\check{s}(n)$ denotes the Hilbert transform of $s(n)$. The envelope, $e(n)$ of the signal is then given by:

$$e(n) = |r(n)|$$

In order to avoid fast fluctuations of the envelope of the signal, the envelope is actually computed as the RMS value of non-overlapped segments of the envelope $e(n)$, where the length of the segment was 2.5 times the mean pitch period of the gender of the speaker (i.e. assuming average fundamental frequency as 120 Hz for male and 200 Hz for female speakers, respectively). Fig. 4 shows an example of the envelope estimation for the signal shown in the lower panel in Fig. 2

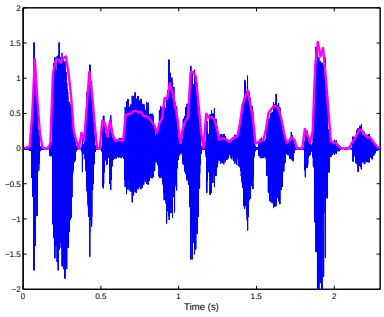

Fig. 4. Example of an envelope estimation. Computed envelope is shown with a thick solid line.

DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the signal is dynamically compressed with 2ms release time constant and has an almost instantaneous attack time constant. More specifically,

$$\hat{e}(n) = \begin{cases} a_r\hat{e}(n-1) + (1-a_r)e(n), & if\ e(n) < \hat{e}(n-1) \\ a_a\hat{e}(n-1) + (1-a_a)e(n), & if\ e(n) \geq \hat{e}(n-1) \end{cases} \quad (5)$$

In the present work, the time constants were selected as $a_r = 0.15$ and $a_a = 0.0001$.

During the static stage the smoothed envelope, $\hat{a}(n)$ is converted to dB and applied to the IOEC curve, shown in Fig. 3, to obtain the time-varying gain. The 0 dB reference level e_0, is a key element in forming the IOEC, is obtained by making an estimate of the largest envelope of the output waveform. For this work, it was set to 30% of the maximum of the envelope of the input signal. Having the reference level, the input envelope is computed in dB

$$e_{in}(n) = 20\log_{10}\left(\hat{e}(n)/e_0\right)$$

The output level $e_{out}(n)$ is obtained from the IOEC curve and then the gain is computed as:

$$g(n) = 10^{(e_{out}(n) - e_{in}(n))/20}$$

The DRC output signal is given by:

$$s_g(n) = g(n)s(n)$$

At the final stage, the global energy of $s_g(n)$ is scaled so that is the same as that of the original unmodified speech signal. An example of the output from DRC is depicted in Fig. 5.

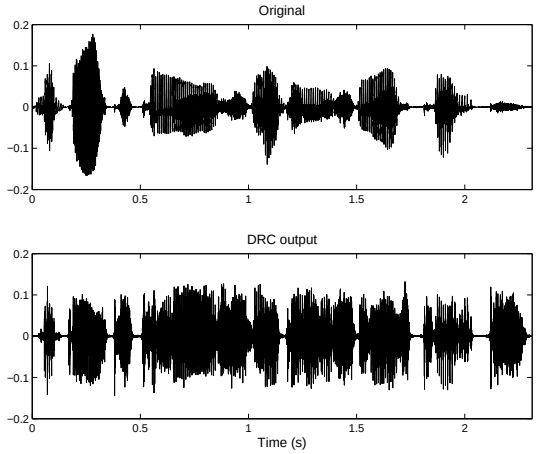

Fig. 5. Example of the output from DRC. Unmodified speech signal (upper panel). Output from DRC (lower panel)

4. RESULTS

For testing the suggested system, we used 240 Harvard sentences uttered by a male speaker, and two types of noise: Speech Shaped Noise (SSN) at SNR: -9dB, -4 dB and 1 dB, and Competing Speaker noise (CS) at SNR: -21 dB, -14 dB, -7 dB. SII was selected to objectively measure the performance of the suggested system and compare it with other published systems. For this purpose the extended SII algorithm was implemented [12] using multi-resolution analysis windows; from $35ms$ for the lowest critical band ($150Hz$) to $9.4ms$ for the highest band ($8000Hz$). Fig. 6 shows the SII scores for the original (unmodified) speech, for the suggested system, and for its subsystems; high pass after frequency reallocation (SRH) and Dynamic Range Compression sub-system (DRC). The cascade combination of these two sub-systems provide the final suggested system(SRHDRC). Both subsystems contribute to the improvement of SII score. DRC seems to perform a bit better than SRH in most of the cases. The combination of these two sub-systems improves the SII score considerably, showing that the two sub-systems are complementary; SRH works in the frequency domain and DRC in the time domain (as it was already mentioned above).

For comparison purposes, the speech-in-noise enhancement systems suggested in [2] and [3] were also implemented and tested on the same Harvard utterances and under the same noise conditions. For all these systems, energy constraints were imposed so that unmodified and modified signal have the same global RMS value. Fig. 7 compares systems SV06

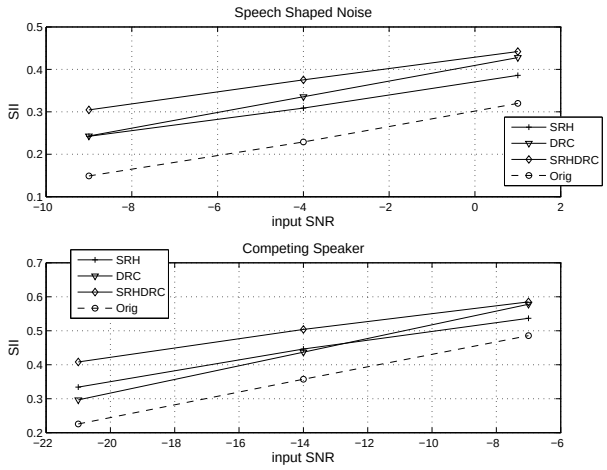

Fig. 6. Speech Intelligibility Index before and after processing with the suggested system for SSN (upper panel) and CS (lower panel).

([2]) and SV10 ([3]) with the suggested system, SRHDRC. Again, the baseline (the SII score for the unmodified speech) is also provided. Overall, we observe that the suggested

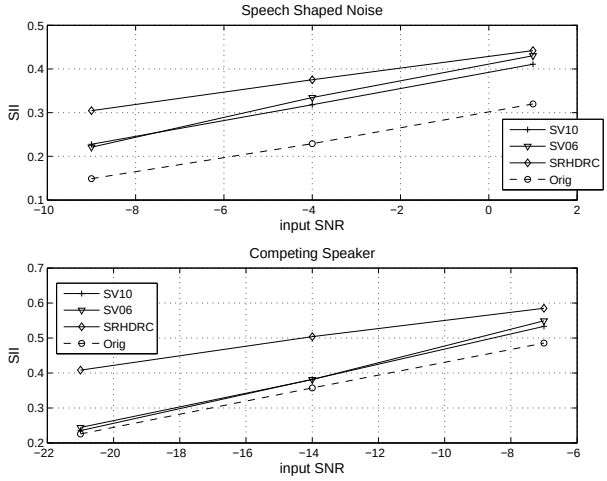

Fig. 7. Speech Intelligibility Index before and after processing with the suggested system and the methods described in [2] (SV06) and [3] (SV10).

system (SRHDRC) outperforms SR for all SNR levels and for both types of noise. All modified signals report better SII score than the non-modified signals. Between SV06 and SV10, the winner is SV06.

Informal listening tests shows that the enhanced speech using the suggested approach indeed produces more intelligible speech than the enhanced signal produced by SV10 or SV06. This is somehow expected since, based on the SII re-

sults shown in Fig. 7, the suggested enhanced system, and for low SNR values (-4dB for SSN, and -14 dB for CS) has an improvement over SV06 and SV10 of 4 dB and 8 dB in terms of SNR gain, for the SSN and the CS types of noise, respectively.

5. CONCLUSIONS

In this work we suggested to enhance the original speech signal combining a signal to noise ratio recovery approach following by a high-pass filtering with dynamic range compression in order to improve the intelligibility of the speech in noise under the constraint of equal signal power before and after the modification. Tests with speech shaped noise and competed speaker noise conditions at various low SNR values, show that the suggested approach outperforms state-of-the art methods in terms of SII score. Moreover the modified signal has not artifacts and actually has a more a crispy quality than the original signal.

6. ACKNOWLEDGMENT

This work was supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623.
This work was produced while Yannis Stylianou was an invited Professor to AhoLab, Univ. of the Basque Country, Bilbao, 2011-2012.

7. REFERENCES

[1] J. Junqua, "The lombard reflex and its role on human listeners," *JASA*, vol. 93, pp. 510–524, 1993.

[2] B. Sauert and P. Vary, "Near end listeniing enhancement: Speech intelligibility improvement in noisy environments," in *Proceedings of IEEE ICASSP-2006*, Toulouse, France, pp. 493–496.

[3] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.

[4] American National Standards Institute, "Methods for the calculation of the speech intelligibility index," *ANSI*, vol. S3.5-1997, 1997.

[5] R.S. Niederjohn and J.H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by highpass filtering followed by rapid amplitude compression," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 277–282, 1976.

[6] V. Hazan and R. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *JASA*, vol. 130, no. 4, pp. 2139–2152, 2011.

[7] R. Smiljanić and A.R. Bradlow, "Production and perception of clear speech in croatian and english," *JASA*, vol. 118, pp. 1677–1688, 2005.

[8] J.C. Krause and L.D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *JASA*, vol. 115, pp. 362–378, 2004.

[9] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble and stationary noise," *JASA*, vol. 124, pp. 3261–3275, 2008.

[10] V. Hazan and A. Simpson, "Cue-enhancement strategies for natural VCV and sentence materials presented in noise," *Speech, Hearing and Language*, vol. 9, pp. 43–55, 1996.

[11] S.D. Yoo, J.R. Boston, A.El-Jaroudi, C.C. Li, J. D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *JASA*, vol. 122, no. 2, pp. 1138–1149, 2007.

[12] 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," *JASA*, vol. 117, no. 4, pp. 2181–2192, 2005.