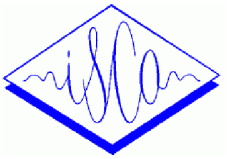
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speech-in-noise intelligibility improvement based on spectral shaping 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

In this paper, we suggest a non-parametric way to improve the intelligibility of speech in noise. The signal is enhanced before presented in a noisy environment, under the constraint of equal global signal power before and after modifications. Two systems are combined in a cascade form to enhance the quality of the signal first in frequency (spectral shaping) and then in time (dynamic range compression). Experiments with speech shaped (SSN) and competing 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). In terms of SNR gain there is an improvement of 7 dB (SSN) and 8 dB (CS) over these methods. A formal listening test confirm the efficiency of the suggested system in enhancing speech intelligibility in noise.

Index Terms: speech-in-noise enhancement, speech intelligibility, spectral shaping, dynamic range compression

1. Introduction

The ability to detect speech in noise plays a significant role in our communication with others. However, 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. Understanding speech in announcements at airports or train stations is also very important. In all these cases, noise may mask part of the speech signal such that not all speech information is available to the listener.

In many intelligibility studies it was demonstrated that clear speech is significantly more intelligible in noise than conversational or casual speech for both normal-hearing and hearing-impaired listeners [1] [2] [3]. Even in a quiet background, clear speech is perceived as more intelligible than casual speech for hearing-impaired listeners and elderly persons as well as for linguistically inexperienced listeners like non-native speakers. 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. This is known as Lombard effect or Lombard reflex [4]. In Lombard speech it has been reported, that there is higher energy in the mid-frequency region of the frequency spectrum, a more spectral flattening (reduced spectral tilt), while no differences in fundamental frequency have been observed [5].

In very early studies for speech-in-noise enhancement, pure signal processing approaches have been tried. Niederjohn and Grotelueschen suggested a rapid amplitude compression following high-pass filtering for processing speech before its reception by the listener [6] [7]. Their work was based on the observation that high-pass filtered/clipped speech offers a significant gain in the intelligibility of speech in white noise over that for unprocessed speech at the same signal-to-noise-ratios (SNR) [6]. More recently, in a series of papers, B. Sauert and P. Vary suggested many speech-in-noise enhancement approaches assuming the noise is known, with and without power constraints (in terms of SNR) [8] [9].

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].

Previous works, they either ignore studies (i.e., [8] [9]) on the acoustic-phonetic features that are enhanced when talkers speak clearly or when they are exposed to noise (Lombard), or they require some type of (broad or not) phonetic classification before modification (i.e., [10], [11]) which, in some cases, it is a serious constraint (i.e., real time speech-in-noise enhancement).

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. The suggested system incorporates results from audio processing and from research in phonetics without, however, the requirement of any type of phonetic classification. The suggested system contains two sub-systems connected in a cascade form. The first sub-system works in the frequency domain and it is an adaptive spectral shaper. Its purpose is to increase the "crisp" and "clean" quality of the speech signal, and therefore improve the intelligibility of speech even in clear (not-noisy) conditions. This is achieved by sharpening the formant information (following observations in clear speech) and by reducing spectral tilt using pre-emphasis filters (following observations in Lombard speech). The specific characteristics of this sub-system are adapted to the degree of speech frame voicing. The second sub-system is inspired by compression strategies used in sound recording and reproduction, audio broadcasting [12] as well in amplification techniques in hearing aids [13]. In this work we reduce the peak to rms value of the speech signal by combin-

This work was partly funded by the E.U. FP7 FET-OPEN (LISTA project: grant agreement 25623) and the Sectoral Operational Program Human Resources Development 2007- 2013 of the Romanian Ministry of Labor, Family and Social Protection through the Financial Agreement POSDRU/88/1.5/S/61178. The work produced while Yannis Stylianou was an invited Professor at the AhoLab, Univ. of the Basque Country, Bilbao, 2011-2012.

ing a downward compression to decrease the loudness of the most sonorant parts of speech with an upward compression to increase the loudness of the less sonorant parts of speech like vocalic onsets and offsets, nasal, stops. Also, we make sure that low energy signals are unaffected. Given the constraint of equal power of the signal before and after modification, the two sub-systems perform re-allocation of energy in frequency (spectral shaping) and in time (dynamic range compression). 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 . Large formal listening test confirmed the effectiveness of the suggested speech-in-noise enhancement approach.

The rest of the paper is organized as follows. In Section 2 we present the energy reallocation algorithm in the frequency domain using spectral shaping. 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 where both objective and subjective results are provided, and finally, Section 5 concludes the paper.

2. Spectral Shaping: Energy reallocation in Frequency

Spectral shaping consists of the following steps. First the probability of voicing of a frame is computed. Next, an adaptive and a fixed spectral shaping operator are applied to the input signal.

2.1. Probability of voicing

To avoid the introduction of artifacts in the processed signal (especially in fricatives, silence or other "quiet" areas of speech) the probability of voicing is used. The probability of voicing is simply defined as:

$$P_v(t) = \alpha \frac{rms(t)}{z(t)} \tag{1}$$

where $\alpha = 1/max(P_v(t))$ is a normalization parameter, $rms(t)$ and $z(t)$ denote the RMS value and the zero-crossings of a segment of the speech signal as this is defined by a rectangular window centered at time t and of length 2.5 the average fundamental period of speaker's gender ($8.3ms$ and $4.5ms$ for males and women, respectively). The probability of voicing is computed frame-by-frame, using a frame rate of 10ms.

2.2. Adaptive spectral shaping

The adaptive spectral shaping consists of (i) adaptive sharpening (for formants enhancement), and (ii) an adaptive pre-emphasis filter. For the enhancement of formant information, a simple approach was followed similar to the one described in [14]. In each frame, and using a Hanning window of the same length as for the computation of the probability of voicing, the N-point DFT of the windowed signal is computed, and using the magnitude spectrum, the spectral envelope of the frame is estimated using SEEVOC [15]. Tilt, $T(\omega)$, of the spectral envelope is computed using cepstrum, as follows:

$$\log T(\omega) = c_0 + 2c_1 \cos(\omega) \tag{2}$$

where

$$c_m = \frac{1}{N/2+1} \sum_{k=0}^{N/2} \log E(\omega_k) \cos(m\omega_k) \tag{3}$$

where $E(\omega_k)$ is the estimated spectral envelope using SEEVOC and c_m is the m^{th} cepstrum coefficients. Then, the first adaptive spectral shaping, at frame t, is defined as:

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)} \right)^{\beta \ P_v(t)} \tag{4}$$

In (4) a formant shaping is performed which is controlled by the local probability of voicing $P_v(t)$ and coefficient β. In this way, no spectral shaping is performed during unvoiced regions, avoiding the creation of tonal sounds. Typical values of β are around 0.25 for low SNR while it can be decreased for higher SNR values.

The next adaptive filter is based on intelligibility experiments made by Niederjohn et al. [6] where a fixed pre-emphasis of 6dB/octave starting at 1100 Hz was shown to improve intelligibility in noise. Such a pre-emphasis filter will introduce a noisy quality to the input speech signal as was also mentioned in [14]. To avoid this, the pre-emphasis filter is not applied on unvoiced areas while the degree of pre-emphasis is controlled by the probability of voicing:

$$H_p(\omega, t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega - \omega_0}{\pi - \omega_0} g \ P_v(t) & \omega > \omega_0 \end{cases} \tag{5}$$

where $\omega_0 = 0.125\pi$, assuming a sampling frequency of $16kHz$. The value g may depend on the SNR level; in this work a constant value (0.3) was used .

2.3. Non-adaptive spectral shaping

The purpose of the fixed (non-adaptive) spectral shaping is to protect the speech signal from low-pass operations during the reproduction of the signal. It is also a pre-emphasis filter which is applied independently of the probability of voicing (since low-pass operation affects all the signal). This filter, $H_r(\omega)$, boosts the energy between 1000 Hz and 4000 Hz by 12 dB, while reduces by 6dB/octave the frequencies below 500 Hz.

The above filters are combined to modify accordingly the magnitude spectrum of the frame:

$$\hat{E}(\omega) = E(\omega) \ H_s(\omega) H_p(\omega) H_r(\omega) \tag{6}$$

The modified amplitude spectrum is combined with the original phase spectrum, and inverse DFT and Overlap-and-Add is used to reconstruct the spectral shaped signal.

3. Dynamic Range Compression: Energy reallocation in Time

The goal of DRC is to produce a time-varying gain to reduce the envelope variations of a signal. This gain is derived from a desired input/output envelope characteristic (IOEC) curve. The type of IOEC used in this work is shown in Fig. 1 with its three characteristic zones; unity gain, expansion, and compression. First, an initial envelope of the speech signal using the analytic signal is computed. To reduce its fast fluctuations over time, the RMS value on non-overlapped segments of the envelope is computed, where the length of the segment is 2.5 times the mean pitch period of speaker's gender ($8.3ms$ and $4.5ms$ for males and women, respectively).
Next step is the compression stage using the estimated envelope, $e(n)$. DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the signal is dynamically

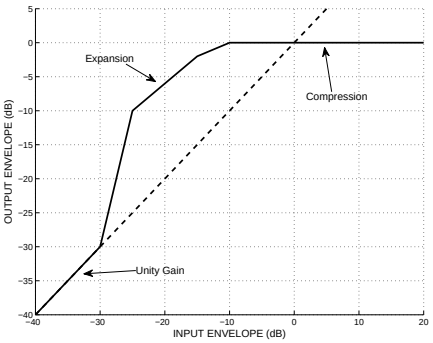


Figure 1: Input-Output Envelope Characteristic (IOEC) curve.

compressed with 2ms release time constant and has an almost instantaneous attack time constant. More specifically,

$$\hat{e}(n) = \begin{cases} a_r\hat{e}(n-1) + (1-a_r)e(n), & if\ e(n) < \hat{e}(n-1) \\ a_a\hat{e}(n-1) + (1-a_a)e(n), & if\ e(n) \geq \hat{e}(n-1) \end{cases} \quad (7)$$

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, $\hat{e}(n)$ is converted to dB and applied to the IOEC curve, shown in Fig. 1, to obtain the time-varying gain. The 0 dB reference level e_0, is a key element in forming the IOEC. In this work, it was set to 30% of the maximum of the input signal envelope. 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. 2.

4. Results

4.1. Objective measurements

For testing the suggested system, we used 240 Harvard sentences uttered by a male speaker (British English), 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. The competing speaker was one female voice. For this purpose the extended SII algorithm was implemented [16] using multi-resolution analysis windows; from $35ms$ for the lowest critical band ($150Hz$) to $9.4ms$ for the highest band ($8000Hz$). For comparison purposes, the speech-in-noise enhancement systems suggested in [8] and [9] 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

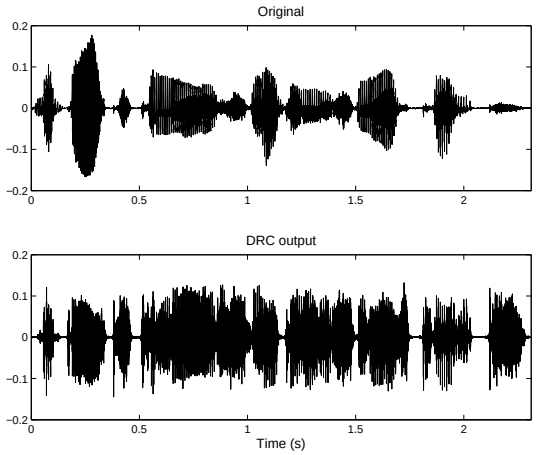


Figure 2: Example of the output from DRC. Unmodified speech signal (upper panel). Output from DRC (lower panel)

signals have the same global RMS value. Fig. 3 shows the SII scores for the original (unmodified) speech, for the suggested system (SSDRC) and for systems SV06 ([8]) and SV10 ([9]). Overall, we observe that the suggested system (SSDRC) outperforms SV06 and SV10 for all SNR levels and for both types of noise. All modified signals report better SII score than the original non-modified signals, for SSN, while for CS, only SSDRC has higher SII score than the unmodified speech. Based on SII scores, SSDRC has an SNR gain over SV06 and SV10 of about 7dB for SSN and 10 dB for CS.

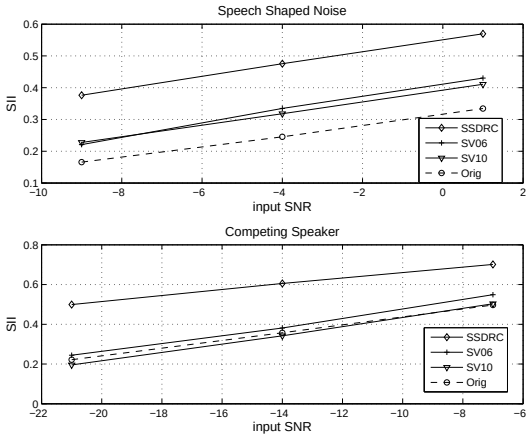


Figure 3: Speech Intelligibility Index before and after processing with the suggested system and our implementations of the methods described in [8] (SV06) and [9] (SV10).

4.2. Formal listening test

A formal listening test has been conducted at the Centre for Speech Technology Research (CSTR) in the University of Edinburgh using the first 180 sentences of the Harvard corpus recorded by a native British English talker. This is part of the same set of data used for the objective measurements described before. In total, 154 listeners with British English as native language with no reported hearing impairment partici-

pated in the listening experiment. The order of sentences and listening conditions was random. All signals were played at 16kHz over headphones to the participants in sound-isolated booths. Listeners never heard the same sentence more than once, and they have been asked to type in what they heard. The type of noise and the corresponding SNR conditions are the same as those used in the objective measurements. Fig. 4 shows the percentage of correct transcription listeners provided for the unmodified and the modified speech using the suggested speech enhancement algorithm (SSDRC), for Speech Shaped Noise (SSN) (upper panel) and Competing Speaker (CS) (lower panel). One female voice was used as the competing speaker. In all cases an SNR gain of 4 dB is obtained by using SSDRC.

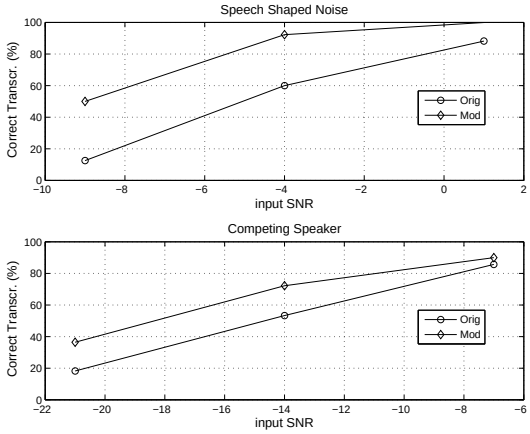


Figure 4: Correct transcription score (%) for original and SSDRC-based modified speech under Speech Shaped Noise (upper panel) and Competing Speaker (lower panel) noisy conditions

5. Conclusions

In this work we suggested to enhance the original speech signal by combining spectral shaping and 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. Objective tests with speech shaped noise and competing 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 is artifacts-free and has a more crispy quality than the original signal. Formal listening test confirms the efficiency of the suggested system in improving speech intelligibility by providing an SNR gain of 4dB for both types of noise.

6. Acknowledgment

Authors would like to thank Cassie Mayo and Vassilis Karaiskos from CSTR/Univ. of Edinburgh and Martin Cooke from Laslab/UPV, for their contribution in the subjective evaluation of the system.

7. References

[1] 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.

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

[3] 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.

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

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

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

[7] R.S. Niederjohn and J.H. Grotelueschen, "Speech intelligibility enhancement in a power generating noise environment," in *Proceedings of IEEE-ICASSP*, 1978, vol. 26, pp. 378–380.

[8] 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.

[9] 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.

[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] B.A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. Audio Acoust.*, vol. 17, no. 1, 1969.

[13] J.M. Kates, "Signal processing for hearing aids," in *In Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandnberg, Eds. Kluwer Acad Pubus: Boston, 1998.

[14] T.F. Quatieri and R. J. McAulay, "Peak-to-rms reduction of speech based on a sinusoidal model," *IEEE Trans. on Signal Processing*, vol. 39, no. 2, pp. 273–288, 1991.

[15] D. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 4, pp. 786–794, 1981.

[16] 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.