NON-LINEAR SOFT-SOUNDS ENHANCEMENT FOR NEAR-END SPEECH INTELLIGIBILITY IMPROVEMENT

Rajyalakshmi Dokku [1,2]

[1] Ruhr-Universität Bochum, Bochum, Germany
[2] FZI Research Center for Information Technology, Karlsruhe, Germany
dokku@fzi.de

ABSTRACT

The objective of this research is to modify the clean speech in a way that it will be more intelligible when it is played in noisy environment without increasing global signal-to-noise ratio. A new near-end speech enhancement algorithm is derived in this contribution based on an extrapolation technique. In this method speech energy is transferred from high energy regions of the speech signal to low energy regions by considering soft-sounds/strong-voiced components classification decisions into account. Variable amplification gain is derived and applied to the classified speech components depending on their original energy levels. The proposed algorithm does not require any information about input noise characteristics for near-end speech enhancement problem.

The derived algorithm is combined with baseline near-end speech enhancement method as a post processing block for testing overall performance. Significant intelligibility improvements are observed with the proposed method over unprocessed noisy speech and considerable improvements are observed with combined method over recent version of the baseline method.

Index Terms— speech detection, speech intelligibility, near-end speech enhancement

1. INTRODUCTION

For a decade near-end speech enhancement problem has been addressed many times in different aspects. When speech communication takes place in noisy environments, speech is less audible to the listener because of the additive noise. Often people require to communicate with others in crowded public places such as at markets and train stations. In such situations normally people change their way of speaking by putting more emphasis on particular phonemes or part of phonemes in order to make their speech more clear and intelligible[1]. In telecommunications, it is always advantageous for the listener if the incoming signal is automatically adjusted to the current situation where the listener is located in order to increase the intelligibility. This is referred to as near-end speech enhancement. Recently many near-end speech enhancement methods were proposed to overcome this problem by assuming input noise characteristics as known. But in reality the noise situations arising in everyday communication are very different from each other and are not predictable perfectly.

In recent years, Sauert and Vary addressed the near-end speech enhancement problem several times and suggested different speech enhancement approaches for improving the intelligibility of speech in noise conditions assuming that the input noise is known[2][3] .

In some of the near-end speech enhancement approaches the power constraint is taken into account for redistributing the speech energy over time and/or frequency as opposed to increasing the play-back level of speech which may not be possible anymore due to loud-speaker limitations or unpleasant play-back levels. Previous research methods also suggest that selective enhancement of vocalic onsets and offsets of speech signal can improve the speech intelligibility for the same overall signal-to-noise ratio[4][5][6].

The average speech spectrum is raised over the average noise spectrum in order to recover a target signal-to-noise-ratio along with the dynamic range compression in [7]. All the above mentioned techniques are mostly developed by assuming input noise conditions as already known and available. In real time applications, the near-end speech enhancement methods which depend on the noise characteristics may not be able to estimate noise floor perfectly unless the additive input noise is also exactly the same as assumed noise. So, improvisation of speech intelligibility is limited for unknown noise conditions.

Speech components like bursts, fricatives, vocalic onsets and offsets are defined as soft-sounds in this paper. A time domain based noise independent selective soft-sounds enhancement algorithm is derived in this contribution for near-end speech enhancement (NSE) applications. Soft-sounds are extracted and enhanced more than strong-voiced sounds in speech signals because they are low in energy and are very important for speech identification and discrimination. In this research an experiment is conducted, in which proposed noise independent near-end speech enhancement method is added to different base-line NSE methods as a post-processing block to observe the possibility of further speech intelligibility improvements.

This paper is organized as follows. Section 2 describes the proposed NSE system (combination of soft-sounds enhancement method with base-line NES method) and then explains derived soft-sounds enhancement algorithm in two steps. First step shows the process for extracting the soft-sounds from continuous speech signals. Second step explains how to derive appropriate amplification gain for each extracted soft speech component. Then, Section 3 presents experimental set-ups and implementation results while section 4 concludes and highlights the main perspectives of this work.

2. PROPOSED NEAR-END SPEECH ENHANCEMENT SYSTEM

In order to improve the near-end speech intelligibility we modify the speech signal both in frequency and time domain. Fig. 1 shows the proposed OLFSSE (optimal linear filtered soft-sounds enhancement) speech modification system which is cascade combination of two sub-systems. The first sub-system is baseline method (optimal linear filter (OLF)) proposed in [8] and the second sub-system is soft-sounds enhancement (SSE) method explained in subsequent

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

sections.

In OLFSSE system speech modification is done by redistributing the speech energy over frequency by using optimal linear filtering as in [8] and then the soft-sounds are detected in processed speech signal by using an extrapolation based detector and enhanced in order to achieve temporal enhancement. The amplification gain is varied for all detected components based on their original signal power. And then the total power of the modified signal is normalized back to the global signal power. The main aim of OLFSSE system is also to see the possibility of speech intelligibility improvement after SII (Speech Intelligibility Index) maximization through the optimal linear filtering in stage-1.

Optimal linear filtering shown in stage-1 of proposed system optimizes the intelligibility of speech in noise for near-end listener by distributing the speech energy over frequency bands such that approximation of the SII is maximized. In stage-2, a new derived soft-sounds enhancement algorithm is applied on every processed utterance in stage-1 to enhance the soft sound frames. The soft-sound speech components detection algorithm will be elucidated in the following sections.

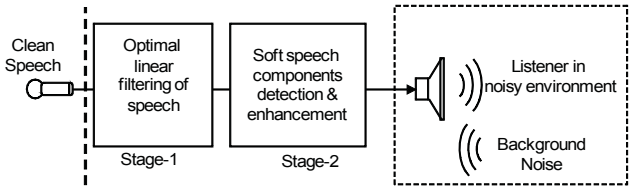

Fig. 1. Proposed OLFSSE system for near-end speech enhancement.

2.1. Soft-Sounds Enhancement (SSE) algorithm

The new soft-sounds detection algorithm is derived based on our previous work in [9]. In our previous work noisy stop consonants detection algorithm was implemented and is modified in this contribution to detect the soft-sounds in clean speech. Extrapolation is a method to extend speech samples in forward direction based on the present observations. In this technique we assume that there exists a set of prediction filter coefficients (a_k) of order p that would linearly predict any sample in a given signal perfectly with zero prediction error i.e.,

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} \tag{1}$$

There are two main factors to describe how soft-sounds detection method is different from the stop consonants detection method published in our previous work. Although we adopt the extrapolation technique from our previous work to detect the soft-sounds, the detection process and detecting components are quite different compared to stop consonants detection method. In our previous stop consonants detection algorithm we used three series of decisions (decision tree) for stop consonants detection with tight thresholds in order to avoid false detections due to the voiced onsets and fricatives. With soft-sounds detection algorithm we want to locate all soft-sounds (as defined in section 1) along with the stop consonants in clean and processed speech. In this contribution we extracted the soft-sound frames based on the energy difference and prediction gain which are computed between original frames and extrapolated frames.

Initially, the clean speech signal S_{plain} is segmented into M frames with frame advance R and frame length N and then the frame based extrapolation technique is implemented. Every extrapolated frame is predicted based on the previous frame of the original speech signal, for more details see [9]. The energy of all original (EO) and extrapolated (EE) frames are computed as follows,

$$EO_i \quad = \quad x_i^T x_i; \quad EE_i = x_{extr_i}^T x_{extr_i} \tag{2}$$

where x_i is the i^{th} speech frame, x_{extr_i} is the i^{th} extrapolated speech frame. The prediction gain and energy difference are used for measuring the success of the prediction. The prediction gain (PG) and the energy difference (ED) are computed as shown in Eq. (3) and (4) respectively,

$$PG_i = 10 \log_{10} \left(\frac{\sigma_{x_i}^2}{\sigma_{e_i}^2} \right) \tag{3}$$

where $\sigma_{x_i}^2$ is the variance of the input clean signal and $\sigma_{e_i}^2$ is the error variance computed between the clean speech frame and corresponding extrapolated frame.

$$ED_i = 10 \log_{10}(EO_i) - 10 \log_{10}(EE_i) \tag{4}$$

where ED_i is the energy difference of the i^{th} speech frame.

Soft sound frames extraction is done based on the following decision in Eq. (5),

$$x_i = \begin{cases} \text{Soft sound frame} & PG_i \leq \tau_{pg} \text{ and } ED_i \geq \tau_{ed} \\ \text{Strong voiced frame} & \text{Otherwise} \end{cases} \tag{5}$$

where τ_{pg} and τ_{ed} are PG and ED thresholds to determine the decision boundaries. As explained with experimental results in [9], prediction gain and energy difference are less during onsets and transients because of sudden rise in energies and noise like behavior in stops. It is difficult to predict the soft-sound signals based on their previous frames by using AR(Auto Regressive) filter model due to the structural and energy difference between them. In order to locate less predictable speech components (like soft-sounds), thresholds of the decision parameters shown in Eq. (5) are computed through experiments for the speech database used in this paper for the experiments. Thus, decision parameters are set to $\tau_{pg} = 0$ dB and $\tau_{ed} = 4$ dB because at these values we observed maximum speech ineligibility improvements over unprocessed speech.

Lets assume I is the vector of size $1 \times M$ which stores the strong-voiced/ soft-sounds frame decisions in terms of $I_i = 0$ for strong-voiced frame and $I_i = 1$ for soft-sound frame. To enhance the detected soft-sound frames based on Eq. (5) we need to derive right enhancement gains for every detected component. Derivation of variable amplification gain for soft-sounds enhancement will be explained in the following section.

2.2. Variable amplification gain derivation

If we assume every speech utterance consists of both strong-voiced and soft-sound components then the total power (P_T) of the speech signal is sum of both the components in that signal. We can formulate it as follows,

$$P_T = P_{SV} + P_{SS} \tag{6}$$

where P_{SV} is the power related to the strong-voiced (SV) components i.e., sum of all SV components power and P_{SS} is the power

related to the soft-sound (SS) components i.e., sum of all SS components power. As SS components are relatively weak sounds and has less energy, amplification is needed to make them audible. Thus, in order to increase intelligibility of speech signal the SS components are amplified with α factor while the total power constraint is achieved by β factor.

$$P_T = \beta^2(P_{SV} + \alpha^2 P_{SS}) \tag{7}$$

By fixing the global amplification gain factor α of complete speech utterance we can compute the value of β by keeping total power constant:

$$\beta^2 = \frac{P_T}{(P_{SV} + \alpha^2 P_{SS})} \tag{8}$$

Since all SS components do not have same properties, applying same amplification gain to all SS components is not a good idea and may not yield good results. For computing the variable gain for each detected soft-sound component the following set of equations are derived.

Let the new amplification gain for every SS frame is denoted by $\hat{\alpha_i}$. The power values (P_{SS}) of the SS frames are extracted as shown in Eq. (9).

$$P_{SS_i} = I_i P_i \tag{9}$$

where P_i is power of the i^{th} frame.

For getting the variable amplification gain, the actual input power of SS frame is subtracted from the maximum value and then normalized by the difference between maximum and minimum values of SS frames in logarithmic domain.

$$\hat{\alpha_i} = \frac{\gamma(\underset{i}{max}(\ln(1 + I_i P_i)) - \ln(1 + I_i P_i))}{\underset{i}{max}(\ln(1 + I_i P_i)) - \underset{i}{min}(\ln(1 + I_i P_i))} + A \tag{10}$$

where γ and A are constant values and are used there to avoid the zero amplification values. The behavior of amplification gain $\hat{\alpha_i}$ is shown in Fig. 2 along with the clean speech signal and energy levels of SS frames. The value of $\hat{\alpha_i}$ is high when the detected soft speech component original energy $I_i P_i$ is low and $\hat{\alpha_i}$ is low when the detected soft speech component original energy $I_i P_i$ is high. As shown in Fig. 2, $\hat{\alpha_i}$ varies according to the detected signal energy even within the phoneme. When the soft-sound component frame has maximum energy, no amplification is applied to that frame because in that case $\hat{\alpha_i} = A$ and A is set to 1 in our experiments.

For simplification, Eq. (10) is modified as follows,

$$\hat{\alpha_i} = (\gamma m_i + A) \tag{11}$$

where m_i is

$$m_i = \frac{(\underset{i}{max}(\ln(1 + I_i P_i)) - \ln(1 + I_i P_i))}{\underset{i}{max}(\ln(1 + I_i P_i)) - \underset{i}{min}(\ln(1 + I_i P_i))} \tag{12}$$

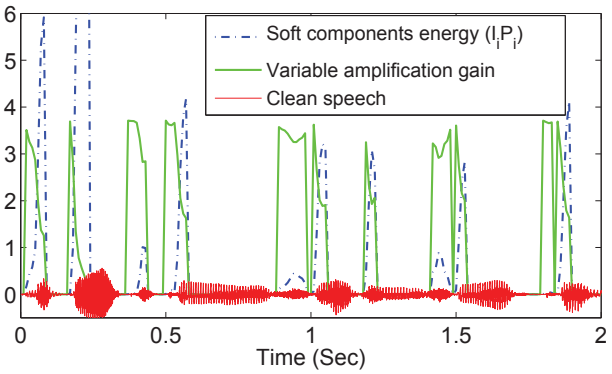

Fig. 2. Variable gain for soft sounds enhancement at global amplification gain $\alpha = 3$.

According to the problem formulation, the sum of the individual powers of the SS frames computed with the new amplification factors $\hat{\alpha_i}$ is equal to the total power of the speech soft-sounds computed with global amplification factor α.

$$\sum_{i=1}^{M} \hat{\alpha_i}^2 I_i P_i = \alpha^2 P_{SS} \tag{13}$$

The value of γ is computed by substituting the Eq. (11) in Eq. (13) ,

$$\sum_{i=1}^{M} (\gamma m_i + A)^2 I_i P_i = \alpha^2 P_{SS}$$

For simplicity $\tilde{P}_{SS}$ and ν_i are used instead of $\alpha^2 P_{SS}$ and $I_i P_i$ respectively in the following derivation.

$$\sum_{i=1}^{M} (\gamma m_i + A)^2 \nu_i = \tilde{P}_{SS}$$

$$\gamma^2 + 2\gamma A \frac{\sum_{i=1}^{M} m_i \nu_i}{\sum_{i=1}^{M} m_i^2 \nu_i} + \frac{A^2 \sum_{i=1}^{M} \nu_i - \tilde{P}_{SS}}{\sum_{i=1}^{M} m_i^2 \nu_i} = 0 \tag{14}$$

γ value is obtained by solving quadratic equation in Eq. (14) . The roots of Eq. (14) are shown in Eq. (15) and (16).

$$\gamma_1 = \frac{-A \sum_{i=1}^{M} m_i \nu_i}{\sum_{i=1}^{M} m_i^2 \nu_i} + \sqrt{\left[-\frac{A \sum_{i=1}^{M} m_i \nu_i}{\sum_{i=1}^{M} m_i^2 \nu_i} \right]^2 - c} \tag{15}$$

$$\gamma_2 = \frac{-A \sum_{i=1}^{M} m_i \nu_i}{\sum_{i=1}^{M} m_i^2 \nu_i} - \sqrt{\left[-\frac{A \sum_{i=1}^{M} m_i \nu_i}{\sum_{i=1}^{M} m_i^2 \nu_i} \right]^2 - c} \tag{16}$$

where $c = \frac{A^2 \sum_{i=1}^{M} \nu_i - \tilde{P}_{UV}}{\sum_{i=1}^{M} m_i^2 \nu_i}$

To ensure that amplification gain is positive and also to avoid information loss by zero amplification, γ value is computed by taking maximum value among two derived roots γ_1, γ_2 and 1. Then the final amplification gain $\hat{\alpha_i}$ for every SS frame is computed by substituting γ value ($max(\gamma_1, \gamma_2, 1)$) in Eq. (11). Modified speech signal is reconstructed by using enhanced SS frames and SV frames.

Output signal power normalization:

Signal power normalization is needed to keep the constraint of global signal power constant before and after speech modifications. After soft-sounds enhancement as explained in above section total power of the modified signal (S_m) is then normalized back to the original signal (S_{plain}) power as shown in Eq. (17).

$$\hat{S}_m = S_m \sqrt{\left(\frac{\sum_{n=1}^{K} S_{plain}^2(n)}{\sum_{n=1}^{K} S_m^2(n)} \right)} \tag{17}$$

where K is length of the speech signal and n is sample time index.

Several experiments are conducted for selecting best global amplification factor α to achieve maximum averaged SII and maximum averaged PESQ (perceptual evaluation of speech quality) values over possible amplification factors (between 1.5 and 4). Among all the amplification factors (1,5 to 4) $\alpha = 3$ shows better SII values with less artifacts. It has been also observed that quality decreases by increasing α value more than 3. Through informal listening we also observed that audibility of soft-sounds and strong-voiced sounds are reasonably balanced at $\alpha = 3$.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed soft-sounds enhancement (PROP-SSE) algorithm derived in Section 2 of this paper was applied on clean speech signals to improve their intelligibility in noise. As shown in Fig. 1, the SSE method was also applied on processed signals by OLF method implemented in [8] and final modified signals are evaluated in experiments with the name PROP-OLFSSE.

The performance of the two proposed algorithms(PROP-SSE and PROP-OLFSSE) were tested and compared with base-line method (Cees-2013) proposed in [8] by using 100 randomly selected speech utterances from the TIMIT database [11] at a sampling rate of 16kHz. All 100 clean speech signals were processed by three methods PROP-SSE, PROP-OLFSSE and Cees-2013 and then mixed with four different noise types car noise, competing speaker noise, white noise and non-stationary noise(recorded at matzo factory)) at SNRs in the range between -20 dB and 20 dB in steps of 5 dB. The performance of proposed algorithms were evaluated in terms of the speech intelligibility index predictions as defined in [10]. The mean SIIs for every algorithm, noise type and SNR are presented in Fig. 3 along with unprocessed noise signals at fs = 16 kHz.

Experimental results in Fig. 3 shows that PROP-SSE method exhibit higher ineligibility values than compared to unprocessed speech for all noise types and SNRs that means by enhancing soft-sounds we can improve the speech ineligibility. The SII results of PROP-OLFSSE method is not only showing that they are better than base-line method in all conditions but also revealing that the speech intelligibility improvements are still possible after SII maximization through baseline method. Another observation is that baseline method performance is not disturbed by adding the post processing method in any noise case and at any SNR level. From these results we can also conclude that the soft-sounds are important for speech intelligibility. Possible reasons for achieving higher SII values with proposed PROP-OLFSSE method compared that of baseline method are that PROP-OLFSSE is independent of input noise characteristics and having both spectral and temporal enhancement features. When near-end speech enhancement methods which depend on input noise characteristics fail to estimate noise floor accurately in any case, adding noise independent post processing method to them will improve the speech intelligibility even at high SNRs and compensates the noise estimation errors.

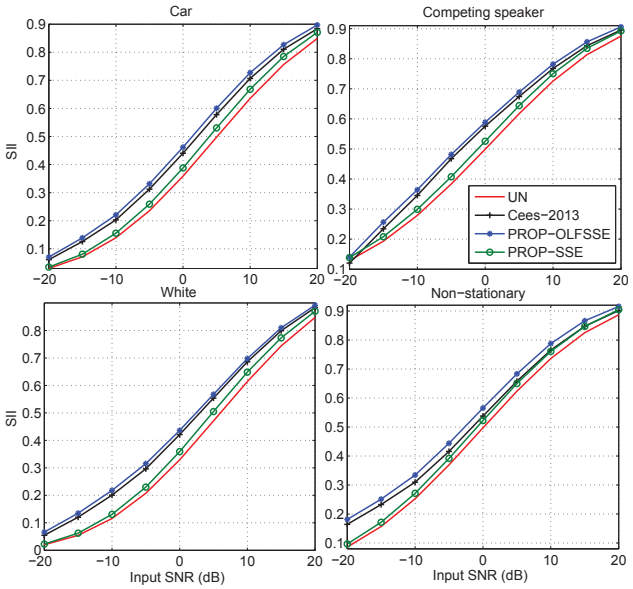

Fig. 3. Speech Intelligibility Index (SII) results for the proposed method (PROP-SSE), combined method (PROP-OLFSSE), baseline method (Cees-2013) and unprocessed noisy speech (UN).

Through informal listening it has been observed that speech understanding in extreme noise conditions with proposed PROP-OLFSSE algorithm is better distinguishable than with other methods discussed in this experiment. In critical applications like near-end speech enhancement, every small SII index improvement is desirable and makes difference in speech understanding for normal hearing listeners.

4. CONCLUSION

In this contribution, a new time-domain based noise independent soft-sounds enhancement algorithm was proposed to improve the intelligibility of speech in noise for the near-end listeners. This was accomplished by selective enhancement of speech components that are important for speech intelligibility. Experimental results of SSE method shows intelligibility improvements over unprocessed noisy speech. Moreover, a new near-end speech enhancement system is proposed by cascading soft-sounds enhancement method with corrected baseline method to achieve speech intelligibility improvements during unknown noise conditions and to compensate noise estimation errors. The proposed OLFSSE system shows better intelligibility improvements over unprocessed noisy speech and considerable improvements over recent near-end speech enhancement method for all SNRs and noise types.

5. REFERENCES

[1] Cooke, Martin, and Maria Luisa García Lecumberri, "The intelligibility of Lombard speech for non-native listeners", The Journal of the Acoustical Society of America ,vol. 132 , pp. 1120-1129, 2012.

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

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

[4] V. Hazan and A. Simpson, "Enhancing information-rich regions of natural vcv and sentence materials presented in noise", in Proceedings of ICASSP, Philadelphia, vol. 1, October 1966, pp. 161-164.

[5] 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", J. Acoust. Soc. Am., vol. 122, no. 2, pp. 1138-1149, 2007.

[6] D. M. Rasetshwane, J. R. Boston, C. C. Li, J. D. Durrant, and G. Genna, "Enhancement of speech intelligibility using transients extracted by wavelet packets," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., New Paltz, NY, USA, 2009, pp. 173–176.

[7] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression," in Proc. EUSIPCO, 2012, pp. 2075–2079

[8] Cees H. Taal, Jesper Jensen, and Arne Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement", IEEE Signal Processing Letters, VOL. 20, NO. 3, MARCH 2013

[9] R. Dokku and R. Martin, "Detection of stop consonants in continuous noisy speech based on an extrapolation technique," in Proc. European Signal Processing Conference (EUSIPCO), Bucharest, Romania, pp. 2338-2342, 2012.

[10] ANSI, "Methods for Calculation of the Speech Intelligibility Index (SII)" ANSI, New York, 1997, S3.5-1997.

[11] J. Garofolo, Timit: Acoustic-Phonetic Continuous Speech Corpus NIST, Boulder, CO, USA, 1993.