SPEECH INTELLIGIBILITY ENHANCEMENT USING TUNABLE EQUALIZATION FILTER

Pinaki Shankar Chanda[1] and *Sungjin Park*[2]

[1]LG Soft India, Bangalore, India E-mail: pinaki.s@lgsoftindia.com
[2]LG Mobile Communication Research Center, Seoul, Korea, E-mail: s_park@lge.com

ABSTRACT

Speech reproduction systems such as mobile handsets are often used in environments with moderate or high level of ambient noise where the intelligibility of the spoken words is degraded heavily. We propose a low-complexity system to increase the intelligibility of far-end clean speech signal to a listener who is located in such environment. An implementation of the proposed system on ARM9 RISC processor in a mobile handset is also reported here. Speech intelligibility index (SII) obtained from the enhanced speech and the subjective test results indicate that the proposed system provides a significant gain in speech intelligibility in environments with moderate or heavy ambient noise with the introduction of minimal processing artifacts.

Index Terms—speech intelligibility, speech processing, acoustic noise

1. INTRODUCTION

We consider enhancing the intelligibility of far-end clean speech signal to a listener who is located in an environment with moderate or high level of noise. To mitigate the problem of degradation of the intelligibility of the spoken words by the ambient noise, a common practice is to increase the power of speech towards a greater signal to noise ratio. However, increasing speech power often causes discomfort and listening fatigue to the listener, particularly when the speech power has to be raised to a favorable signal-to-noise ratio in presence of heavy ambient noise. Speech intelligibility enhancement, in such scenarios, calls for enhancement of the perceptual features of the spoken words that are associated with speech intelligibility.

Intelligibility of the spoken words is generally associated with the formant structure of speech signal. In [1] the experimental results indicate that the first formant alone is a very minor contributor to the intelligibility of speech, whereas a strong correlation is observed between the intelligibility of speech and the second formant frequency. It is also known that the consonants play more significant role compared to the vowels in carrying the speech intelligibility cues even though the consonants are significantly weaker than vowels in phonetic power. As consonants carry less phonetic power, they are more prone to be affected by noise when speech is reproduced in an environment with moderate or high ambient noise level. The different frequency bands in speech contribute different amounts to the intelligibility of the spoken words. Frequency range from 1.5 KHz to 3.5 KHz has more contribution in the intelligibility of the spoken words compared to the rest of the speech spectrum. Refer [2] for relative contributions of different frequency bands in speech intelligibility.

In literature different speech intelligibility enhancement techniques have been proposed that are based on the enhancement of the perceptual cues that are associated with the intelligibility of the spoken words. An amplitude compressor followed by a high-pass filter, having a pre-determined cut-off frequency, is used in [3] to enhance the intelligibility of input speech signal. In [4] speech intelligibility is enhanced by equalizing the consonants to the vowels; the spectral contents of speech that are generally associated with the consonants are emphasized by multiplying them with a set of weights. In [5] it is noted that the voiced speech having non-uniform periodicity tends to have lower intelligibility and the speech enhancement is achieved by providing uniform periodic characteristic to voiced speech segments. In [6]-[7] the formant frequencies are emphasized compared to the rest of the frequencies to obtain speech with increased intelligibility. In [8] input speech is enhanced by filtering using a transfer function that approximates the inverse of the Fletcher-Munson curves (the transfer functions of the human hearing system). In [9] a variable cut-off frequency high pass filter has been used to filter the input speech to increase its intelligibility. In [10] the transient components of speech are selectively amplified and recombined with original speech for increased intelligibility.

Many speech intelligibility enhancement systems, proposed earlier, involve high computational and storage complexity that often preclude their implementation on resource limited embedded platforms such as mobile handsets. This can be particularly attributed to the speech enhancement systems that are based on analysis of the spectral components of input speech in frequency domain and selective enhancement of a subset of these frequency components. It is also desirable that while increasing intelligibility of speech the enhancement system should preserve the clarity of the input speech. The intelligibility should be enhanced with minimal introduction of audible processing artifacts. This paper proposes a low-complexity approach to enhance the intelligibility of speech. The proposed system is amenable for implementation in a resource-constrained embedded platform. In this proposal, the consonants of the speech signal are enhanced by processing the input speech using a tunable band-pass shelving filter whose cut-off frequency is dynamically adjusted. The proposed system preserves the speech clarity well without introducing audible distortions. We observe that excess sibilant levels are sometime produced by the proposed system due the boost in the high frequency region of the unvoiced fricatives. To mitigate this problem the proposed system employs a vocal de-esser in conjunction with the speech enhancement unit.

Rest of this paper is organized as follows. In section 2 the proposed speech enhancement system is described. Section 3 reports the results of an integer point implementation in a mobile

handset. In section 4 we present the simulation results and section V concludes this paper.

2. SPEECH INTELLIGBILITY ENHANCEMENT SYSTEM

A block diagram of the proposed system is shown in Figure 1. Input speech is filtered by a high pass shelving filter whose cut-off frequency is adjusted such that the level of the output speech is approximately equal to the level of input speech. The shelving filter is having a gain greater than unity in the high frequency range whereas in the low frequency range the gain of the shelving filter is less than unity.

During a transition from a consonant to a vowel the cut-off frequency of the shelving filter is initially set at a high frequency value as consonants carry dominant high frequency components. While making a transition from a consonant to a vowel, output speech level estimate becomes lower compared to the input speech level estimate as vowels carry dominant low-frequency components and the cut-off frequency of the shelving filter is set at a higher frequency range. The cut-off frequency of the shelving filter is thus shifted to a low frequency range and the shift is proportional to the difference between the input speech level estimate and output speech level estimate to maintain the equality between the input speech level and output speech level. As phonetic powers of the vowels are significantly greater than the phonetic powers of the consonants, the input level estimator output and output level estimator output start rising from a lower level towards a higher level when a vowel appears after a consonant. The attack time of input level estimator and the attack time of the output level estimator are kept equal and they are significantly smaller than their respective release times; hence the system readily shifts the cut-off frequency of the shelving filter towards the low-frequency range to attain the equality between the input and output speech levels without raising the phonetic power of the vowels.

On the other hand when a consonant appears after a vowel in a speech segment the cut-off frequency of the shelving filter is set at a low frequency value. Hence the output speech level estimate becomes greater than the input speech level estimate as the consonants carry dominant high frequency contents. The system shifts the cut-off frequency to a higher frequency range to maintain the equality between input speech level and output speech level. As the phonetic power of the consonant is significantly less compared to the preceding vowel the input level estimator output as well as the output level estimator output starts releasing their estimated levels from a higher value towards a lower value. The release time of output level estimator is smaller than the release time of the input level estimator, and hence in the initial duration of the consonants the cut-off frequency is not shifted momentarily to the higher frequency range. This results in initial boost of the phonetic power of the consonants. The cut-off frequency of the tunable shelving filter is also changed in such a way that there is an upper-limit beyond which the cut-off frequency is not moved even if the output speech level is greater than the input speech level. As a result in case of most of the consonants the equality between the input speech level and output speech level is not satisfied and the consonants get a boost in their phonetic power. A low-pass second order shelving filter with cut-off frequency of 6 KHz is used after the tunable shelving filter to reduce the excess boost of high frequency components beyond its cutoff frequency.

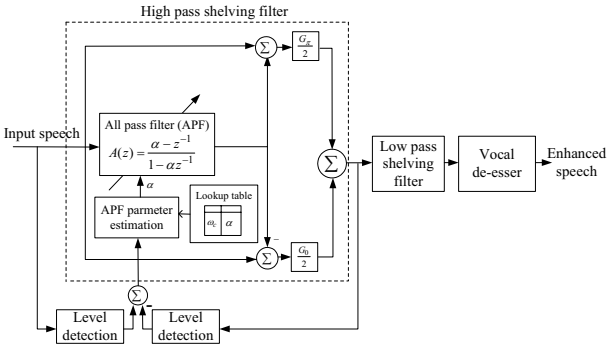

Figure 1. The proposed speech intelligibility enhancement system

From experimental results we observe that the proposed speech enhancement system sometime produces excess sibilant levels. A sibilant is the "ess" vocal sound that is generated while producing unvoiced fricatives such as the 's' in 'say'. A sibilant is characterized by its predominantly high frequency content that has a sharp amplitude peak. Most of the energies of the sibilants vocals are located above 2 KHz. To mitigate the problem with excess sibilant level a vocal de-esser [10] is used along with the speech enhancement unit. In the vocal de-esser the input speech is split into high frequency components and low frequency components using a second-order low-pass shelving filter with cut-off frequency around 2 KHz. When the ratio of the R.M.S. level of the high-pass band frequency content and the R.M.S. level of the low-pass band frequency content exceeds a certain threshold a decision is taken in favor of a sibilant and the de-esser gain is reduced from unity to a lower value using a predetermined release time constant. When non-sibilant vocal appears at the input of the de-esser the gain is increased towards unity with a predetermined attack time constant.

3. REAL-TIME IMPLEMENTATION

The proposed speech intelligibility enhancement system is implemented on integer-point ARM926-EJS processor in a mobile handset as a postprocessor to the vocoder. The tunable high pass shelving filter $H_{\mathrm{SH}}(z)$ is realized using an all-pass filter $A(z)$, leading to a low-sensitivity realization robust to the coefficient quantization [12]. It is shown in Figure 1. $H_{\mathrm{SH}}(z)$ has a gain G_0 at zero frequency and a gain G_π at high frequency range.

$$H_{\mathrm{SH}}(z)=\frac{G_\pi}{2}\left(1+A(z)\right)+\frac{G_0}{2}\left(1-A(z)\right) \tag{1}$$

$$A(z)=\frac{\alpha-z^{-1}}{1-\alpha z^{-1}} \tag{2}$$

The 3-dB cut-off frequency of the filter ω_c is given by

$$\omega_c = \cos^{-1}\left(\frac{2\alpha}{1+\alpha^2}\right) \qquad (3)$$

In this implementation the cut-off frequency and the filter gain can be changed independently of each other permitting easy tuning of the system. For different values of ω_c, the corresponding values of the all-pass filter parameter α are stored in a lookup table. The estimated cut-off frequency of the shelving filter is proportional to the difference between the output speech level and input speech level. Depending on the values of the estimated target cut-off frequency, the all-pass-filter coefficient α is obtained from the lookup table to tune the shelving filter. The parameters G_0 and G_π are experimentally determined. The speech enhancement unit and the de-esser use first-order filters to estimate the input and output speech levels. Table 1 summarizes the computational amount and memory requirement of the proposed speech enhancement system including the vocal de-esser for 8 KHz speech input.

TABLE I. COMPUTATIONAL AMOUNT AND MEMORY REQUIREMENT

Millions of cycles/second (MCPs)	Code	Read-only data	Read-write data
1.20	772 bytes	702 bytes	100 bytes

4. SIMULATION RESULTS

Figure 2 illustrates the effects of the intelligibility enhancement process on a speech segment */people affected by flood-waters of hurricane Katrina/* of a male speaker in frequency-domain. The enhanced speech spectrogram shows more prominent second and higher formant frequencies in the vowels as well as enhanced consonants. In an environment with background noise, the background noise masks a part of the speech spectrum such that only a fraction of the intelligibility cues are available to the listener. Speech Intelligibility Index (SII) in ANSI S3.5-1997 standard quantifies the amount of available speech intelligibility in presence of background noise [3]. It is a physical measure that is highly correlated with the intelligibility of speech. The ratio of time-averaged speech power and time-averaged noise power is measured in a set of frequency bands and SII is computed by adding up the speech-to-noise ratios after multiplying them with a set of weights; the weights correspond to the contribution of the frequency bands in determining speech intelligibility. Good communication systems have an SII of 0.75 or above, while poor communication systems have an SII below 0.45. Figure 3 shows the improvement in speech intelligibility index for speech segments of a male speaker and a female speaker for different signal-to-noise ratios. The noise recordings are obtained from NOISEX-92 database [13]. The speech segments are scaled to 65 dBA. The sampling frequency is 19 KHz. We notice significant improvement in speech intelligibility in case of the male speaker whereas for the female speaker the improvement is moderate in high signal-to-noise ratios.

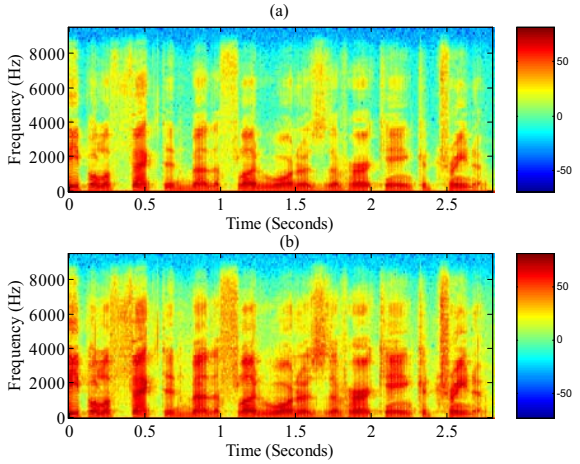

Fig 2. (a) Spectrogram of a speech segment from a male speaker (b) spectrogram of the speech segment after speech intelligibility enhancement

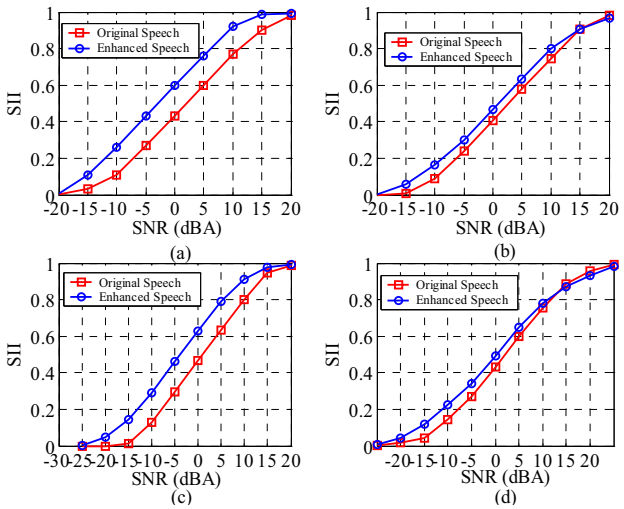

Fig 3. Enhancement in SII using the proposed system. (a) Speech from a male speaker embedded in factory noise (b) Speech from a female speaker embedded in factory noise (c) Speech from a male speaker embedded in babble noise (d) Speech from a female speaker embedded in babble noise

To analyze the improvement of speech intelligibility index on individual vowels and consonants we use a short-term measure of speech intelligibility. The SII computation procedure in ANSI S3.5-1997 standard is extended in [14] by a short-term speech intelligibility measure to capture the effect of fluctuation of ambient noise and to predict speech intelligibility in individual phoneme level. In [14] SII is computed within small time windows as a short-term measure. The sampling frequency of input speech is 19 KHz. Every 9.4 milliseconds the short-term SII is obtained using the speech-to-noise power ratios in 21 frequency bands spanning 19 KHz. Figure 4 shows short term SII improvement in a speech clip of a male speaker in presence of car interior noise obtained from NOISEX-92 database. From Figure 4 we observe the relative increase in SII is more in consonants whereas for the vowels the SII is not enhanced significantly. Figure 5 shows the improvement in time-averaged short-term SII over the duration

two speech segments from a male speaker and a female speaker for different signal–to-noise ratios in presence of factory noise and babble noise from NOISEX-92 database. From Figure 3 and Figure 5 it is evident that the proposed system substantially boosts the speech intelligibility index in presence of moderate and high level of background noise.

TABLE II. COMPARISON SCALE

Scale	Meaning	Scale	Meaning
3	Much better	-1	Slightly worse
2	Better	-2	Worse
1	Slightly better	-3	Much worse
0	About the same		

Informal listening tests are carried out to characterize the performance of the speech enhancement system. Six listeners are asked to rate the intelligibility of the original speech clips embedded in background noise compared to the intelligibility of the enhanced speech clips embedded in the same background noise on a 7 point comparison scale given in Table II . This scale is according to ITU-T P.830 recommendations [15]. The speech clips are from male and female speakers. Factory noise recordings at 3 different signal-to-noise ratios {-5 dB, 0 dB, and 5 dB} are used for the test. The mean ratings obtained from the listeners are {1.52 1.45 1.44} respectively for the three SNRs.

5. CONCLUDING REMARKS

In this paper we have proposed a low-complexity system for improving intelligibility of speech in presence of moderate or high level of ambient noise. From simulation results and informal listening tests it is observed that the proposal is quite effective in enhancing speech intelligibility in adverse environments. In ongoing work we are evaluating the improvement of speech intelligibility using formal subjective tests.

6. REFERENCES

[1] Ian B. Thomas, "The influence of first and second formants on the intelligibility of clipped speech," Journal of the Audio Engineering Society, vol. 16, no. 2, pp. 182–185, Apr. 1968.

[2] American National Standard, "Methods for the Calculation of the Speech Intelligibility Index," ANSI S3.5-1997, 1997.

[3] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Trans. Acoust., Speech, and Sig. Proc., vol. 24, no. 4, pp. 277 – 282, Aug 1976

[4] J.M. Kates, "Speech intelligibility enhancement," U.S. Patent 4454609, June 12, 1984.

[5] J.M. Kates and J.J. Bussgang, "Speech enhancement techniques," U.S. Patent 4468804, 1984

[6] T. Mekata, "Formant detecting device and speech processing apparatus," U.S. Patent 5479560, 1995

[7] A.L. Klayman, "Public address intelligibility system," U.S. Patent 5459813, 1995

[8] A.L. Klayman, "Voice intelligibility enhancement system," U.S. Patent 6993480, 2006

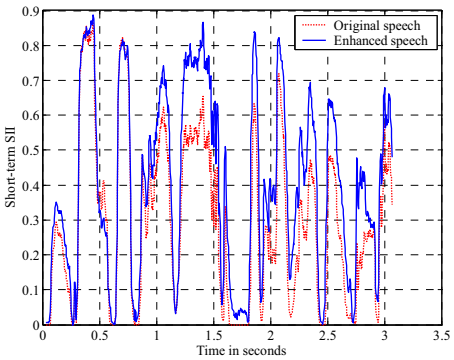

Fig 4. Improvement in short-term SII of a male speech segment in presence of car interior noise. The speech segment is scaled to 65 dBA and noise recording is scaled to 60 dBA

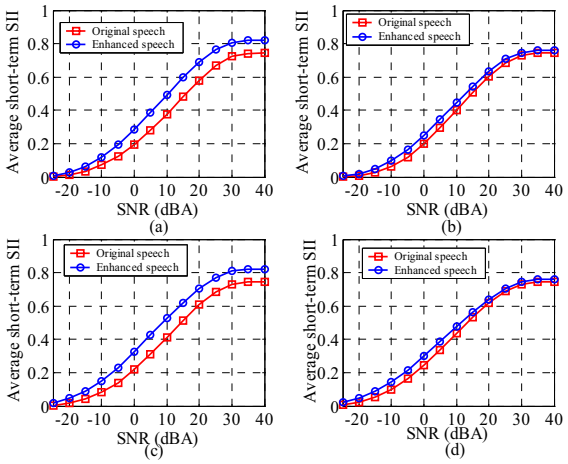

Fig 5. Enhancement in time-averaged short-term SII using the proposed system. (a) Speech from a male speaker in factory noise (b) Speech from a female speaker in factory noise (c) Speech from a male speaker in babble noise (d) Speech from a female speaker in babble noise

[9] M. Vierthaler, "Circuit for improving the intelligibility of audio signals containing speech," U.S. Patent application 20020173950, 2002

[10] C. Tantibundhit, J.R. Boston, C.C. Li, J.D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "Speech enhancement using transient speech components," In proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1 pp. 833-836, 2006

[11] J.B. Lemanski, "A new vocal de-esser,"69th AES. Conv., May 1981

[12] P. Regalia and S. Mitra, "Tunable digital frequency response equalization filters," IEEE Trans. Acoust., Speech, and Sig. Proc., vol. 35, no. 1, pp. 118 – 120, Jan 1987

[13] NOISEX-92 database-http://spib.rice.edu/spib/select_noise.html

[14] 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," Jour. Acoustic Soc. Amer., 117(4), Pt. 1, April 2005.

[15] ITU-T P.830, Telephone transmission quality methods for objective and subjective assessment of quality, ITU-T Recommendation, 1996.