A Novel Speech Intelligibility Improvement Method Using Maximizing Mutual Information Measure

Elham Eideli
Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran.

Seyed Mohammad Ahadi
Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran.

Neda Faraji
Department of Electrical Engineering, Imam Khomeini International University, Qazvin, Iran.

Summary
In many public areas, such as air or rail terminals, Public Address (PA) systems play a significant role in announcing important messages. However, the intelligibility of messages announced by such public address systems decreases due to environmental noise sources. A speech pre-processing algorithm is proposed for speech intelligibility improvement in these noisy environments. The proposed algorithm modifies the original clean speech such that it would be more intelligible for the listener in the presence of additive background noise. Our proposed algorithm maximizes the mutual information between the temporal envelopes of the clean and noisy modified speech in sub-band domain for energy redistribution of clean signal under an energy constraint. Evaluations using two objective intelligibility measures show that our proposed algorithm provides significant gain over the unprocessed speech signal and has higher scores in comparison with the reference method.

PACS no. 43.71.-k, 43.71.Gv

1. Introduction

One of the most important goals in speech communication systems such as Public Address (PA) system is conveying message in a way that would be comprehensible for the listeners, i.e. has high intelligibility. PA systems are involved in many institutional and commercial buildings and locations like airports, train stations, sports stadiums and many other large public spaces. In this paper, we consider a simple PA system as depicted in Figure 1. Unfortunately, different environmental issues like background additive noise and reverberation which are also perceived alongside the announcement, degrade speech intelligibility at the listener side.

Conventional speech enhancement methods are not applicable in PA systems, as they typically operate on an additive mixture of speech and noise signals. Furthermore, they are mostly proposed to improve the quality of speech and are not capable of enhancing speech intelligibility, and even in some cases they possibly deteriorate it [1]. In the intelligibility improvement application, as the listeners are located in a noisy environment, the noise signal reaches their ears without hardly any possibility to intercept. Hence, the only practical approach to increase the intelligibility is to manipulate the clean speech in a pre-processing unit before making announcement [2].

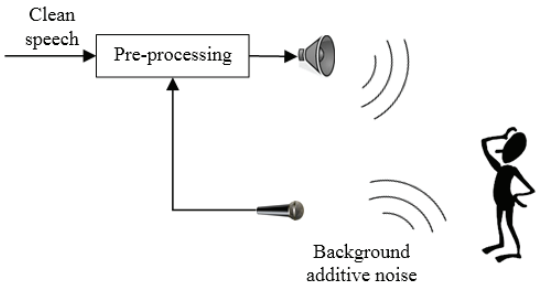

Figure 1. Example of a Public Address system.

Various speech modifications exist for processing clean speech signal in order to make it more intelligible in the presence of background noise [2]-[5]. Some of these modifications are inspired from strategies that are normally adopted by speakers in the noisy environments [6],[7]. For instance, the so called Lombard speech produced in noisy

environments has high intelligibility in comparison with speech produced in quite [8]. However, it has shown that some of the changes embedded in the Lombard speech have no significant contribution to speech intelligibility, such as increasing pitch, and are probably emerged from the constraint of speech production mechanism [6]. Totally, speech modification algorithms could be classified in two distinct categories. The first-category methods modify clean signal by enhancing perceptual cues that are important in intelligibility of speech [6],[7]. The other category algorithms modify clean signal using an optimization procedure. These algorithms modify clean signal by, for example, energy redistribution over time, frequency or in time-frequency regions to optimize an intelligibility measure. In these algorithms, optimal modifications depend on both clean and noise signals [3],[5],[10]. In this paper a new speech intelligibility measure is suggested for improving speech intelligibility in the presence of background additive noise. Recently, different objective speech intelligibility measures based on Mutual Information (MI) have been proposed for intelligibility prediction of speech signal corrupted by different types of linear and non-linear distortions [11],[12]. In this paper, we utilize an MI-based measure for optimization purpose. In our proposed algorithm, the energy of clean speech is redistributed in a way that maximizes the MI between temporal envelopes of clean and processed noisy speech signals.

The remainder of this paper is organized as follows. First, in Sec. 2 the mutual information and MI-based intelligibility measure are described. Then, the MI-based speech modification approach for improving intelligibility is given in Sec. 3. The experimental results and conclusion are explained in Sec. 4 and 5, respectively.

2. Mutual Information and Intelligibility Measure

Mutual information is a non-parametric measure of relevance that measures the mutual dependence of two variables, either linear or non-linear [9]. MI between two discrete random variables X and Y is defined as [9]

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) ln\left(\frac{p(x,y)}{p(x)p(y)}\right), \quad (1)$$

where $p(x,y)$ is the joint and $p(x)$ and $p(y)$ are the marginal probability distribution functions of the variables.

Mutual information can be equivalently rewritten by entropy values

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \quad (2)$$

where $H(X,Y)$ is the joint and $H(X)$ and $H(Y)$ are the marginal entropies [9]. For the case of Gaussian random variables, the mutual information is defined as [9]

$$I(X;Y) = -\frac{1}{2}\ln(1-\rho^2), \quad (3)$$

where ρ is the linear correlation coefficient between X and Y. MI gets minimum value of zero in case of two independent random variables. The more dependency, the higher MI value is obtained. In this paper, we utilize equation 3 to define an intelligibility measure for speech modifications.

3. Speech Modification Approach Using MI-Based Intelligibility Measure

In this work, improving speech intelligibility is achieved by redistributing the energy of clean speech signal over both time and frequency under an energy constraint. The energy constraint guarantees that the energy of speech signal remains constant before and after the modification. This is due to this fact that in some applications it is more desired to enhance intelligibility without any increase in global energy level of speech signal [6]. The block diagram of our proposed system is depicted in Figure 2. Spectral modification (energy redistribution in different frequency bands) is similar to the reference system [10] and temporal modification (energy redistribution in time) is carried out by Dynamic Range Compression (DRC) method [14].

3.1. Spectral modification

Spectral modification is carried out during an utterance by optimizing our suggested MI-based intelligibility measure. The suggested measure is computed as follows: first, Time-Frequency (TF) representations of both clean and processed noisy signals are obtained by framing, windowing and applying DFT to each frame. Then, DFT-bins are split into some non-overlapping frequency bands that are linearly spaced on a Mel scale. Assume $\bar{y}$, x and $\tilde{x}(k,l)$ represent time-domain noisy processed signal, clean speech and the k^{th} DFT-bin of l^{th} short-time frame of clean speech signal, respectively. For l^{th} frame of clean speech the energy of j^{th} sub-band is computed by

$$X_j(l) = \sqrt{\sum_{k\epsilon K_j}|\tilde{x}(k,l)|^2}, \quad (4)$$

where K_j indicates all DFT-bins which belong to j^{th} frequency band. The TF representation for $\bar{y}$ is obtained similarly, and will be represented by $\bar{Y}_j$. The intelligibility measure is defined in sub-band domain as in [11] that is during the utterance MI is computed for each frequency band using equation 3 and then these values are averaged for all the frequency bands as

$$\bar{I} = \frac{1}{J}\sum_{j=1}^{J} I(X_j; \bar{Y}_j), \qquad (5)$$

where $\bar{I}$ and J represent the average MI value of the utterance and the total number of frequency bands, respectively. In this paper, equation 5 is the intelligibility measure suggested to enhance speech intelligibility.

Using TF representation of clean speech and modification parameters obtained in the optimization procedure, clean signal is modified and synthesized. For evaluating intelligibility measure TF representation of both clean and noisy processed signals are required. A realization of noise signal is then added to the clean modified speech in order to obtain an estimation of noisy modified speech. Optimization process continues until reaches a local optimum. Optimal parameters which are gains for adjusting energy of each frequency band are then used to modify clean speech.

3.2. Temporal modifications

Temporal modification in [10] is applied using gain adjustment of phone energies which needs each phoneme boundaries obtained by forced alignment algorithm using acoustic models from an Automatic Speech Recognition (ASR) system. Each phone's energy is adjusted using a single parameter under the energy constraint. In applications such as PA systems in which this information is not available, and consequently the suggested measure could not be used for this purpose, temporal modification could effectively completed by DRC technique. DRC is a successful technique [14] in improving speech intelligibility. In [14] spectral shaping is combined with DRC technique. Reducing spectral tilt which is done by spectral shaping has positive impact on intelligibility of speech. However, modifying spectral tilt depends on noise spectral profile and is not a good strategy in the presence of high frequency noise [6]. Therefore, combining spectral modification using MI-based intelligibility measure with DRC would be beneficial, because in this case spectral tilt modification depends on noise characteristic. DRC reduces the envelope variations of the signal, and effectively redistributes signal's energy in time domain. DRC in this work is based on explanation in [14].

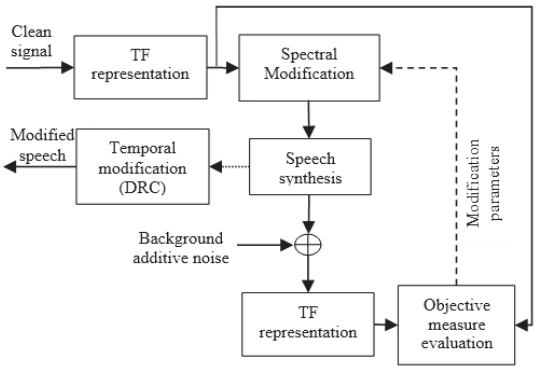

Figure 2. The block diagram of the proposed system.

The reference system [10] uses a high level intelligibility measure for spectral and temporal modifications. It needs acoustic models from an ASR system, and the state-level alignment. Our suggested system utilizing MI-based intelligibility measure has no need to this high-level information, and so has low complexity.

A sample speech signal and its spectrogram which is modified by our suggested system for babble noise at the SNR level of -15 dB are depicted in Figure 2 and Figure 4, respectively. Natural (unprocessed) signal and modified signal using suggested system are shown in these figures. As depicted, after modification energy is transferred from energetic regions (e.g. vowels and low frequency regions) to low energy regions (e.g. consonants and middle and high frequency regions).

4. Experimental Results

For testing the suggested system, we used 50 Harvard sentences (set 40-44) uttered by a male speaker at 16000 Hz sample rate, and four different noise types: babble, factory, train and white noises at SNR: -15 dB, -10 dB, -5 dB, 0 dB and 5 dB. The frame length and frame shift used for computing MI-based measure are 25 ms and 10 ms, respectively. Two objective intelligibility measures are used to measure the performance of the proposed system and compare it with reference system [10]. CSII [15] and GP [16] are two objective intelligibility measures which have shown high correlation with the subjective intelligibility

scores [15],[16]. CSII is an extended version of SII which is suitable for measuring non-linear distortions [15]. GP measures the audibility of speech in the presence of noise [17]. CSII output ranges from zero to one, and GP output ranges from zero to 100. The higher score in both GP and CSII, the higher intelligibility would be.

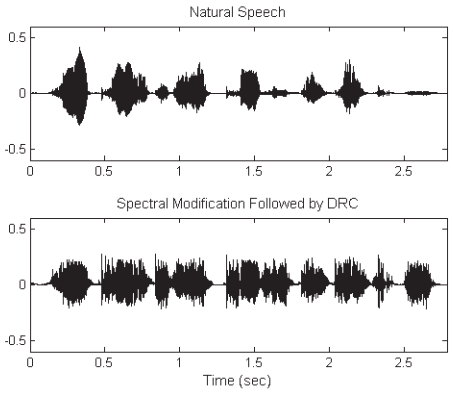

Figure 3. A sample signal modified by suggested system. Natural (unmodified) signal (top), modified signal using MI-based measure and DRC (bottom).

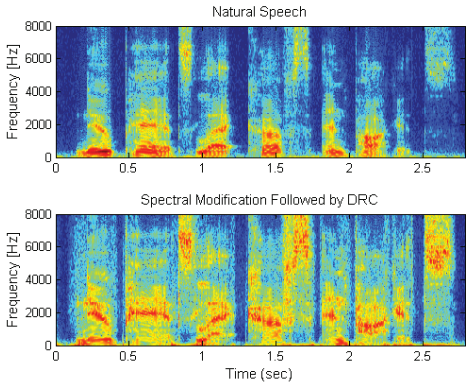

Figure 4. Spectrogram of the sample signal modified by suggested system. Natural (unmodified) signal (top), modified signal using MI-based measure and DRC (bottom).

For comparison purposes, the suggested system in [10] is used as a reference method and tested on the same utterances and noisy conditions. In Figure 5 and Figure 6, CSII and GP scores of the original (unprocessed) speech and modified speech signals using proposed and reference systems are depicted. Using CSII scores, we could say that our suggested system outperforms the reference system especially for babble, factory and train noises. GP scores validate the CSII scores except in the white noise case where intelligibility improvement has been obtained only for high SNRs. In the low SNRs of white noise, both reference and proposed system show no noticeable intelligibility improvement in terms of both GP and CSII scores. This result is in line with the fact that white noise, as a stationary noise has more negative impact on speech intelligibility [17].

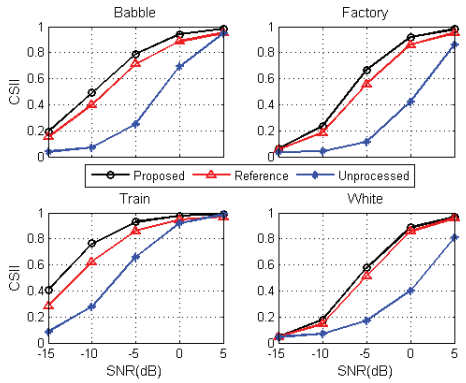

Figure 5. CSII intelligibility score before (unprocessed) and after processing with the suggested system and the method described in [10].

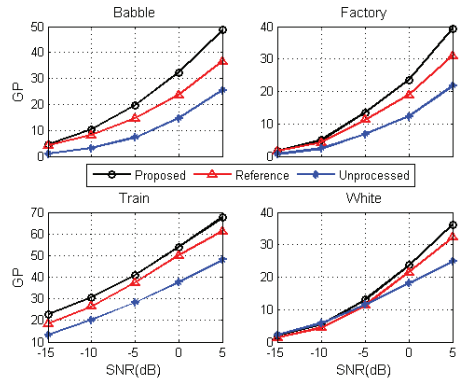

Figure 6. GP intelligibility score before (unprocessed) and after processing with the suggested system and the method described in [10].

5. Conclusions

A novel objective intelligibility measure based on mutual information is suggested to improve speech intelligibility in the presence of background additive noise. Using this intelligibility measure, the energy of clean speech signal is redistributed over frequency under the energy constraint. For the temporal energy redistribution DRC technique is utilized. Performance evaluations using two objective intelligibility measures under different

noisy conditions reveal that the proposed system, though its lower complexity, outperforms the reference system.

References

[1] P. C. Loizou and G. Kim: Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Transactions on Audio, Speech, and Language Processing 19 (2011) 47-56.

[2] B. Sauert and P. Vary: Near end listening enhancement: Speech intelligibility improvement in noisy environments. Proc. ICASSP 2006, 493-496.

[3] C. H. Taal, R. C. Hendriks, and R. Heusdens: A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. Proc. ICASSP 2012, 4061-4064.

[4] Y. Tang and M. Cooke: Energy reallocation strategies for speech enhancement in known noise conditions. Proc. INTERSPEECH 2010, 1636-1639.

[5] V. Aubanel and M. Cooke: Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. Proc. INTERSPEECH 2013, 3592-3596.

[6] M. Cooke, S. King, M. Garnier, and V. Aubanel: The listening talker: A review of human and algorithmic context-induced modifications of speech. Computer Speech & Language 28 (2014) 543-571.

[7] D. Rasetshwane, J. Boston, J. Durrant, S. Yoo, C.-C. Li, and S. Shaiman: Speech enhancement by combination of transient emphasis and noise cancelation. In Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (2011), 116-121.

[8] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes: Effects of noise on speech production: Acoustic and perceptual analyses. The Journal of the Acoustical Society of America 84 (1988) 917-928.

[9] T. M. Cover and J. A. Thomas: Elements of information theory. John Wiley & Sons, 2012.

[10] P. N. Petkov, G. E. Henter, and W. B. Kleijn: Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise. IEEE Transactions on Audio, Speech, and Language Processing 21 (2013) 1035-1045.

[11] J. Taghia and R. Martin: Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing. IEEE/ACM Transactions on Audio, Speech and Language Processing 22 (2014) 6-16.

[12] J. Jensen and C. H. Taal: Speech intelligibility prediction based on mutual information. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (2014) 430-440.

[13] Y. Lu and M. Cooke: Speech production modifications produced in the presence of low-pass and high-pass filtered noise. The Journal of the Acoustical Society of America 126 (2009) 1495-1499.

[14] T-C. Zorila, V. Kandia, and Y. Stylianou: Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[15] J. M. Kates and K. H. Arehart: Coherence and the speech intelligibility index. The Journal of the Acoustical Society of America 117 (2005) 2224-2237.

[16] M. Cooke: A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America 119 (2006) 1562-1573.

[17] M. Cooke: Glimpsing speech. Journal of Phonetics 31 (2003) 579-58.