WHY DO SPEECH-ENHANCEMENT ALGORITHMS NOT IMPROVE SPEECH INTELLIGIBILITY?

Gibak Kim and Philipos C. Loizou

Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA
Email: imkgb27@gmail.com, loizou@utdallas.edu

ABSTRACT

While most speech enhancement algorithms improve speech quality, they do not improve speech intelligibility in noise. The reasons for that remain unclear. In this paper, we present a theoretical framework that can be used to analyze potential factors influencing the intelligibility of processed speech. It is hypothesized that if distortions are properly controlled, then large gains in intelligibility can be achieved. To assess the perceptual effect of the various distortions that can be introduced by speech enhancement algorithms, intelligibility tests are conducted with human listeners. The results indicated that certain distortions are more critical than others. The result of listening tests suggested that when these distortions are properly controlled, substantial gains in intelligibility can be obtained.

Index Terms— Speech intelligibility, speech distortion, speech enhancement

1. INTRODUCTION

While large advances have been reported to suppress background noise and improve speech quality in noise [1, 2], little progress has been made in improving speech intelligibility [3]. The development of an algorithm that would improve speech intelligibility in noisy environments has been elusive for several decades. Little is known as to why speech enhancement algorithms do not improve speech intelligibility. In this paper, we discuss factors responsible for the absence of intelligibility improvement with existing speech enhancement algorithms. The majority of these factors center around the fact that none of the existing algorithms are designed to improve speech intelligibility, as they utilize a cost function that does not necessarily correlate with speech intelligibility. The statistical-model based algorithms (e.g. MMSE, Wiener filter), for instance, derive the magnitude spectra by minimizing the mean-squared error (MSE) between the clean and estimated (magnitude or power) spectra (e.g., [4]). The MSE metric, however, pays no attention to positive or negative differences between the clean and estimated spectra. A positive difference between the clean and estimated spectra would suggest attenuation distortion, while a negative spectral difference would suggest amplification distortion. The perceptual effect of these two distortions on speech intelligibility cannot be assumed to be equivalent. In this paper, we show analytically that if we can somehow manage or control these two types of distortions, then we should expect to receive large gains in intelligibility. To further support our hypothesis, intelligibility listening tests are conducted with normal-hearing listeners.

2. CONSTRAINTS ON THE ESTIMATED MAGNITUDE SPECTRA

To investigate the impact of the two distortions[1] (attenuation and amplification) on speech intelligibility, we use an objective measure which has been found to be highly correlated ($r = 0.81$) with speech intelligibility [6]. More precisely, the measure based on the signal-to-residual spectrum ratio at frequency bin k is given by

$$SNR_{\mathrm{ESI}}(k) = \frac{X^2(k)}{\left(X(k) - \hat{X}(k)\right)^2} \tag{1}$$

where $X(k)$ denotes the clean magnitude spectrum and $\hat{X}(k)$ denotes the magnitude spectrum *estimated* by a speech enhancement algorithm. Dividing both numerator and denominator by $D^2(k)$, where $D(k)$ denotes the noise magnitude spectrum, we get:

$$SNR_{\mathrm{ESI}}(k) = \frac{SNR(k)}{\left(\sqrt{SNR(k)} - \sqrt{SNR_{\mathrm{ENH}}(k)}\right)^2} \tag{2}$$

where $SNR(k) \triangleq X^2(k)/D^2(k)$ is the true *a priori* SNR at bin k, and $SNR_{\mathrm{ENH}}(k) \triangleq \hat{X}^2(k)/D^2(k)$ is the enhanced SNR (note that this is not the same as the output SNR). Based on Eq. (1), we can divide the speech distortions into multiple regions:

[1] Only magnitude-spectrum distortion is considered here, since the phase distortion has not been found to be important in the context of speech enhancement [5].

Region I. In this region, $\hat{X}(k) \leq X(k)$, suggesting only attenuation distortion.

Region II. In this region, $X(k) < \hat{X}(k) \leq 2 \cdot X(k)$ suggesting amplification distortion up to 6.02 dB.

Region III. In this region, $\hat{X}(k) > 2 \cdot X(k)$ suggesting amplification distortion of 6.02 dB or greater.

From the above, we can deduce that in the union of Regions I and II, which we denote as Region I+II, we have the following constraint:

$$\hat{X}(k) \leq 2 \cdot X(k). \quad (3)$$

The constraint in Region I stems from the fact that in this region, $SNR_{\text{ENH}}(k) \leq SNR(k)$ leading to $\hat{X}(k) \leq X(k)$. The constraint in Region II stems from the fact that in this region $SNR(k) < SNR_{\text{ENH}}(k) \leq SNR(k) + 6.02$ dB. Finally, the condition in Region III stems from the fact that in this region $SNR_{\text{ENH}}(k) > SNR(k) + 6.02$ dB. It is clear from the above definitions of these three regions that in order to maximize SNR_{ESI} (and consequently maximize speech intelligibility), the estimated magnitude spectra $\hat{X}(k)$ need to be contained in regions I and II (note that the trivial, but not useful, solution that maximizes $SNR_{\text{ESI}}(k)$ is $\hat{X}(k) = X(k)$). Intelligibility listening tests were conducted to test this hypothesis. If the hypothesis holds, then we expect to see large improvements in intelligibility.

3. INTELLIGIBILITY LISTENING TESTS

5.1. Signal processing

The noise-corrupted sentences were processed by Wiener algorithm based on *a priori* SNR estimation [7]. Let $Y(k,t)$ denote the magnitude of the noisy spectrum at time frame t and frequency bin k. Then, the estimate of the signal spectrum magnitude is obtained by multiplying $Y(k,t)$ with a gain function $G(k,t)$ as follows:

$$\hat{X}(k,t) = G(k,t) \cdot Y(k,t) \quad (4)$$

The Wiener gain function is is given by:

$$G_{\text{Wiener}}(k,t) = \sqrt{\frac{SNR_{prio}(k,t)}{1 + SNR_{prio}(k,t)}} \quad (5)$$

where SNR_{prio} is the *a priori* SNR estimated using the decision-directed approach as follows:

$$SNR_{prio}(k,t) = \alpha \cdot \frac{\hat{X}^2(k,t-1)}{\hat{D}^2(k,t-1)} + (1-\alpha) \cdot \max\left[\frac{Y^2(k,t)}{\hat{D}^2(k,t)} - 1, 0 \right] \quad (6)$$

where $\hat{D}(k,t)$ is the estimate of the background noise spectrum magnitude and α is a smoothing constant (typically set to $\alpha = 0.98$). The noise estimation algorithm proposed in [8] was used for estimating the noise spectrum in Eq. (6).

Oracle experiments were run in order to assess the full potential on speech intelligibility when the proposed constraints were implemented. The magnitude spectrum of the clean speech signal was assumed to be known. The various constraints were implemented as follows. The noisy speech signal was first segmented into 20 ms frames (with 50% overlap between frames), and then processed through the Wiener algorithm, producing at each frame the estimated magnitude spectrum $\hat{X}(k)$. The estimated magnitude spectrum $\hat{X}(k)$ was compared against the true spectrum $X(k)$, and spectrum components satisfying a given constraint were retained, while spectral components violating the constraints were zeroed-out. For the implementation of the Region I constraint, for instance, the modified magnitude spectrum $X_M(k)$, was computed as follows:

$$X_M(k) = \begin{cases} \hat{X}(k) & \text{if } \hat{X}(k) < X(k) \\ 0 & \text{else} \end{cases} \quad (7)$$

An IFFT was finally taken of $X_M(k)$ (using the noisy speech signal's phase spectrum) to reconstruct the time-domain signal. The overlap-and-add technique was subsequently used to synthesize the signal. As shown in Eq. (7), the constraints are implemented by applying a binary mask to the *estimated* magnitude spectrum.

Fig. 1 shows example spectrograms of signals synthesized using the Region I constraints (panel d). The original signal was corrupted with babble at -5 dB SNR. The Wiener algorithm was used in this example, and speech processed with the Wiener algorithm is shown in panel c. As can be seen, the signal processed using the Region I constraints resembles the clean signal, with the formants and voiced/unvoiced boundaries clearly preserved.

5.2. Methods and procedure

Seven normal-hearing listeners participated in the listening experiments, and all listeners were paid for their participation. The listeners participated in a total of 12 conditions (= 2 SNR levels (-5 dB, 0 dB) × 6 processing conditions). For each SNR level, the processing conditions included speech processed using: Wiener algorithm imposed with (1) no constraints imposed, (2) Region I constraints, (3) Region II constraints, (4) Region I+II constraints, and (5) Region III constraints. For comparative purposes, subjects were also presented with noise-corrupted ((6) unprocessed) stimuli.

The listening experiment was performed in a sound-proof room (Acoustic Systems, Inc) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Prior to the sentence test, each subject listened to a set of noise-corrupted sentences to be familiarized with the testing procedure. During the test, subjects were asked to write down the words they heard. Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The order of the conditions was randomized across subjects. The testing session lasted for about 2 hrs. Five-minute breaks were given to the subjects every 30 minutes.

Sentences taken from the IEEE database [9] were used as test material. The sentences in the IEEE database are phonetically balanced with relatively low word-context predictability. The sentences (approximately 2.5 secs. in duration) were originally recorded at a sampling rate of 25 kHz and downsampled to 8 kHz. Noisy speech was generated by adding babble noise at 0 dB and -5 dB SNR. The babble noise was constructed using 20 talkers with equal number of female and male talkers. To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified intermediate reference system (IRS) filters used in ITU-T P.862 [10]. Telephone speech was used as it is considered particularly challenging (in terms of intelligibility) owing to its limited bandwidth (4 kHz). Consequently, we did not expect the performance to be limited by ceiling effects.

5.3. Results

Fig. 2 shows the results of the listening tests expressed in terms of the percentage of words identified correctly by normal-hearing listeners. The bars indicated as "UN" show the scores obtained with noise-corrupted (un-processed) stimuli. As shown in Fig. 2, performance improved dramatically when the Region I constraints were imposed. Performance at -5 dB SNR improved from 5 % correct when no constraints were imposed, to 90 % correct when Region I constraints were imposed. Performance degraded to near zero when Region III constraints were imposed.

Statistical tests, based on Fisher's LSD test, were run to assess significant differences between the scores obtained in the various constraint conditions. Performance of the Wiener algorithm with Region I constraints did not differ statistically ($p>0.05$) from performance obtained with the Region I+II constraints. This was found to be true for both SNR levels. Performance obtained with no constraints did not differ significantly ($p>0.05$) from performance obtained with unprocessed (noise corrupted) sentences for both SNR levels tested. In summary, the analysis indicates that the Region I and Region I+II constraints are robust yielding consistently large benefits in intelligibility for both SNR levels (-5, 0 dB).

4. DISCUSSION AND CONCLUSIONS

Current enhancement algorithms can improve speech quality but not speech intelligibility [3]. The quality and intelligibility are two of the many attributes (or dimensions) of speech and the two are not necessarily equivalent. As shown in [2, 3], algorithms that improve speech quality do not improve speech intelligibility.

The findings of the present study suggest two interrelated reasons for the absence of intelligibility improvement with existing speech enhancement (SE) algorithms. First, and foremost, SE algorithms do not pay attention to the two types of distortions introduced when applying the suppression function to noisy speech spectra. Both distortions are treated equally in most SE algorithms, since the MSE metric is used in the derivation of most suppression functions (e.g., [4]). As demonstrated in Fig. 2, however, the perceptual effects of the two distortions on speech intelligibility are not equal. Of the two types of distortion, the amplification distortion (in excess of 6 dB) was found to bear the most detrimental effect on speech intelligibility (see Fig. 2). Performance dropped near zero when stimuli were constrained in region III. Theoretically, we believe that this is so because this type of distortion (region III) leads to negative values of SNR_{ESI}. In contrast, the attenuation distortion (region I) was found to yield the least effect on intelligibility. In fact, when the region I constraint was imposed, large gains in intelligibility were realized. Performance at -5 dB SNR, improved from 5% correct with stimuli enhanced with the Wiener algorithm to 90% correct when region I constraint was imposed. Theoretically, we believe that the improvement in intelligibility is due to the fact that region I always ensures that $SNR_{\mathrm{ESI}} \geq 0$ dB. Maximizing SNR_{ESI} ought to maximize intelligibility, given the high correlation of a weighted-version of SNR_{ESI} (termed fwSNRseg in [6]) with speech intelligibility. Hence, by imposing the appropriate constraints (see Eq. (3)), we can ensure that $SNR_{ESI} \geq 0$ dB, and subsequently obtain large gains in intelligibility.

Second, none of the existing SE algorithms was designed to maximize a metric that correlates highly with intelligibility. The only known metric, which is widely used in CASA, is the ideal binary mask (IdBM) criterion. This metric maximizes the articulation index (AI), an index that is known to correlate highly with speech intelligibility [11]. Hence, it is not surprising that speech synthesized based on the IdBM criterion improves intelligibility [12]. In fact, it restores speech intelligibility to the level attained in quiet (near 100% correct) even for sentences corrupted by

background noise at SNR levels as low as -10 dB SNR [12]. The IdBM criterion is a special case of the proposed constraint in region I+II, when no suppression function is applied to the noisy spectra, i.e., when $\hat{X}(k) = Y(k)$. From a practical point of view, the proposed constraints are easier to implement than the IdBM criterion, as they do not require access to the noise spectra.

In summary, in order for SE algorithms to improve speech intelligibility they need to treat the two types of distortions differently. More specifically, SE algorithms need to be designed so as to minimize the amplification distortions. As the data in Fig. 2 demonstrated, SE algorithms can improve speech intelligibility if the amplification distortions are properly controlled. Alternatively, and perhaps, equivalently, SE algorithms need to be designed so as to maximize a metric (e.g., SNR_{ESI}, AI) that is known to correlate highly with speech intelligibility. For instance, SE algorithms need to be designed to maximize SNR_{ESI} rather than minimize the un-constrained MSE cost function, as done by most statistical-model based algorithms (e.g., [4]). Algorithms that maximize the SNR_{ESI} metric are likely to provide substantial gains in intelligibility.

ACKNOWLEDGEMENT

This research was supported by Grant No. R01 DC007527 from National Institute of Deafness and other Communication Disorders, NIH.

11. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Florida, 2007.

[2] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement," Speech Communication, vol. 49, pp. 588-601, 2007.

[3] Y. Hu and P. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am*., vol. 22, no. 3, pp. 1777-1786, 2007.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.

[5] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. ASSP*-30, no. 4, pp. 679-681, 1982.

[6] J. Ma, Y. Hu and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387-3405, 2009.

[7] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 629-632, 1996.

[8] S. Rangachari and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," *Speech Communication*, vol. 28, pp. 220-231, 2006.

[9] IEEE Subcommittee, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, no. 3, pp. 225-246, 1969.

[10] ITU, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation P. 862*, 2000.

[11] K. Kryter, "Validation of the articulation index," *J. Acoust. Soc. Am.,* vol. 34, pp. 1698-1706, 1962.

[12] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.,* vol. 123, no. 3, pp. 1673-1682, 2008.

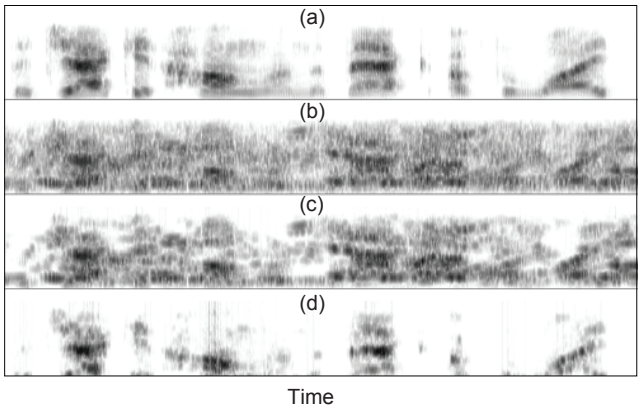

Fig. 1. Wide-band spectrograms of the clean signal (panel a), noisy signal in -5 dB SNR babble (panel b), signal processed by the Wiener algorithm (panel c), and signal processed by the Wiener algorithm after imposing the constraints in Region I (panel d).

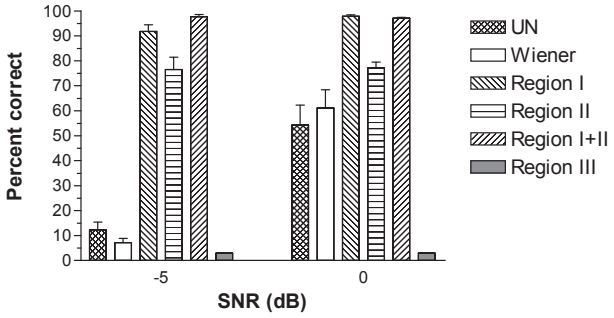

Fig. 2. Results, expressed in percentage of words identified correctly, from the intelligibility studies with human listeners. The bars indicated as "UN" show the scores obtained with noise-corrupted (un-processed) stimuli, while the bars indicated as "Wiener" show the baseline scores obtained with the Wiener algorithm (no constraints imposed). The intelligibility scores obtained with speech processed by the Wiener algorithm after imposing the four different constraints are labeled accordingly.