A COMPARATIVE INTELLIGIBILITY STUDY OF SPEECH ENHANCEMENT ALGORITHMS

Yi Hu and Philipos C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX, USA.
{yihuyxy, loizou}@utdallas.edu

ABSTRACT

In this paper, we report on the evaluation of intelligibility of speech enhancement algorithms. IEEE sentences were corrupted by four types of noise including babble, car, street and train at two SNR levels (0dB and 5dB), and then processed by eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical model based and Wiener-type algorithms. The processed speech files were presented to normal hearing listeners for identification in formal listening tests. Intelligibility was assessed as the percentage of words identified correctly. This paper reports the results of the intelligibility tests.

Index Terms— Speech enhancement, speech intelligibility, speech quality, subjective listening test.

1. INTRODUCTION

The objective of speech enhancement algorithms is to improve one or more perceptual aspects of noisy speech, most notably, quality and intelligibility. Improving quality, however, might not necessarily lead to improvement in intelligibility. In fact, in some cases improvement in quality might be accompanied by a decrease in intelligibility. This is due to the distortion imparted on the clean speech signal resulting from excessive suppression of acoustic noise.

In some applications, the main goal of speech enhancement algorithms is to improve speech quality while preserving, at the very least, speech intelligibility. Hence, much of the focus of most speech enhancement algorithms has been to improve speech quality. Only a small number of algorithms have been evaluated using formal intelligibility tests [1–4], and in those studies, only a single speech enhancement algorithm was evaluated and in a limited number of noise conditions. It therefore remains unclear as to which of the many speech enhancement algorithms proposed in the literature performs well in terms of speech intelligibility. At the very least, we would like to know which algorithm(s) preserve or maintain speech intelligibility in reference to the noisy (unprocessed) speech, and which algorithm(s) impair speech intelligibility, particularly in extremely low SNR conditions. Given the absence of accurate and reliable objective measure to predict the intelligibility of speech processed by enhancement algorithms, we must resort to formal listening tests to answer the above questions.

In this paper, we report on the intelligibility evaluation of eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. This is a follow-up study on the subjective speech quality comparison reported in [5]. Phonetically-balanced sentences were corrupted by four different types of noise commonly encountered in daily life, and processed by the above enhancement algorithms. The enhanced speech files were presented to normal-hearing subjects in a double-walled sound-proof booth, and asked to identify the words in the spoken sentences. This paper presents the summary of these intelligibility tests.

2. TESTING PROCEDURE

IEEE sentences [6] were used in the listening tests. The IEEE database was selected as it contains phonetically-balanced sentences with relatively low word-context predictability. The IEEE sentences were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences, produced by one male speaker, were originally sampled at 25 kHz and downsampled to 8 kHz. 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 [7] for evaluation of the PESQ measure.

Noise was artificially added to the sentences as follows. The IRS filter was independently applied to the clean and noise signals to bandlimit the signals to 3.2kHz. The active speech level of the filtered clean speech signal was first determined using the method B of ITU-T P.56. A noise segment of

Research is supported in part by Grant No. R01 DC07527 from NIDCD/NIH.

the same length as the speech signal was randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal. The noise signals were taken from the AURORA database [8] and included the following recordings from different places: babble, car, street, and train. The noise signals were added to the speech signals at SNRs of 0 dB and 5 dB.

The noise-corrupted sentences were processed by eight different speech enhancement algorithms which included: the generalized KLT approach [9], the perceptual KLT approach (pKLT) [10], the Log Minimum Mean Square Error (logMMSE) algorithm [11], the logMMSE algorithm with speech presence uncertainty (logMMSE-SPU) [12], the spectral subtraction algorithm based on reduced delay convolution (RDC) [13], the multiband spectral subtraction algorithm (MB) [14], the Wiener filtering algorithm based on wavelet-thresholded (WT) multitaper spectra [15], and the Wiener algorithm based on a-priori SNR estimation (Wiener-as) [16]. Detailed description of the eight algorithms can be found in [5]. Matlab implementations of all algorithms are available in [17]. A total of 24 native speakers of American English were recruited for the listening tests. The subjects were paid for their participation. The 24 listeners were divided into four panels (one per type of noise) each consisting of six listeners. Each panel of listeners listened to sentences corrupted by a different type of noise. This was done to ensure that no subject listened to the same sentence twice. Each subject participated in a total of 19 listening conditions (= 2 SNR levels ×8 algorithms + 2 noisy references + 1 quiet). Two sentence lists (10 sentences per list) were used for each condition. The presentation order of the listening conditions was randomized among subjects. The processed speech files, along with the clean and noisy speech files, were presented monaurally to the listeners in a double-walled sound-proof booth via Sennheiser's (HD 250 Linear II) circumaural headphones at a comfortable level. Tests were conducted in multiple sessions with each session lasting no more than two hours. The subjects were allowed to take break during the listening session to reduce fatigue.

3. EVALUATION RESULTS

Listening tasks involved sentence recognition in noise. Speech intelligibility was assessed in terms of percentage of words identified correctly. All words were considered in the scoring. Fig. 1 shows the mean intelligibility scores for babble and car noises, and Fig. 2 shows the mean scores for street and train noises. The error bars in the figures give the standard errors of the mean. The intelligibility scores of noisy (unprocessed) speech are also given for comparative purposes. The intelligibility of sentences corrupted by babble noise was found to be the lowest compared to the other types of noise.

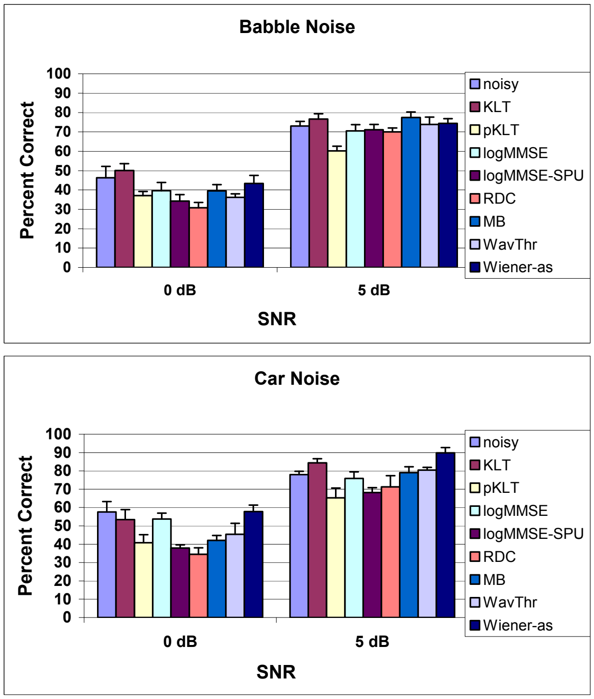

Fig. 1. Mean intelligibility scores of the eight speech enhancement algorithms for the babble and car noise conditions at 0 dB and 5 dB.

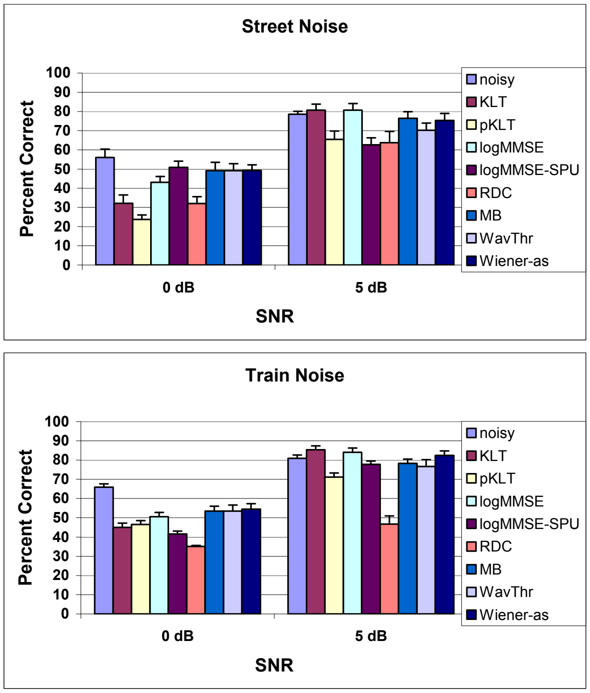

Fig. 2. Mean intelligibility scores of the eight speech enhancement algorithms for the street and train noise conditions at 0 dB and 5 dB.

4. STATISTICAL ANALYSIS

We present comparative analysis at two levels. At the first level, we compare the performance of the various algorithms across all classes aiming to find the algorithm(s) that performed the best across all noise conditions. At the second level, we compare the performance of all algorithms in reference to the noisy speech (unprocessed). This latter comparison will provide valuable information as to which, if any, algorithm(s) improve significantly the intelligibility of noisy speech. If no improvement is obtained, we can learn at the very least which algorithm(s) maintain speech intelligibility and which algorithm(s) diminish speech intelligibility.

In order to assess significant differences between the intelligibility scores obtained from each algorithm, we subjected the scores of the 24 listeners to statistical analysis. Analysis of variance (ANOVA) indicated a highly significant effect (F(8,40)=3.8, $p < 0.005$) of speech enhancement algorithms on speech intelligibility (a highly significant effect was found in all SNR conditions and types of noise). Following the ANOVA, we conducted multiple comparison statistical tests according to Fisher's LSD test to assess significant differences between algorithms. Differences between scores were deemed significant if the obtained p value (level of significance) was smaller than 0.05.

4.1. Intelligibility comparison among algorithms

As shown in Figures 1 and 2, the difference in performance among algorithms was more evident in the 0 dB SNR condition than in the 5 dB SNR condition. At 5 dB SNR, most algorithms performed equally well. At 0 dB SNR, the KLT, logMMSE and Wiener-as algorithms performed equally well for most conditions. In babble noise (0 dB SNR), the KLT and Wiener-as algorithms performed the best among all algorithms. In car noise (0 dB SNR), the KLT, logMMSE and Wiener-as algorithms performed equally well, and significantly better than the other algorithms. At 5dB SNR, five algorithms (KLT, logMMSE, MB, WT and Wiener-as) performed equally well in most conditions. Considering all conditions, the Wiener-as algorithm performed consistently well in all conditions, followed by the KLT and logMMSE algorithms which performed well in six of the eight noise conditions, followed by the WT and MB algorithms which performed well in five and four conditions respectively.

4.2. Intelligibility comparison against noisy speech

Further analysis was performed to find out whether intelligibility is improved or at least maintained (i.e., speech was equally intelligible) in reference to noisy (unprocessed) speech. Statistical analysis revealed that five algorithms (KLT, logMMSE, MB, WT, and Wiener-as) maintained speech intelligibility in six of the eight noise conditions tested. That is, enhanced speech was found to be as intelligible as that of noisy (unprocessed) speech. In one condition (car noise, 5 dB SNR), the Wiener-as algorithm improved the intelligibility of speech. All algorithms produced a decrement in intelligibility of speech corrupted by train noise at 0 dB SNR. The pKLT and RDC algorithms reduced significantly the intelligibility of speech in most conditions.

5. CONCLUSIONS

This paper compared the intelligibility of speech produced by eight different enhancement algorithms operating in several types of noise and SNR conditions. Based on the statistical analysis, we can draw the following conclusions:

1. With the exception of a single noise condition (car noise at 5dB SNR), no algorithm produced significant improvements in speech intelligibility. The majority of the algorithms (KLT, logMMSE, MB, WT, Wiener-as) tested were able to maintain intelligibility at the same level as that of noisy speech.

2. When comparing the performance of the various algorithms, we found that the Wiener-as algorithm performed consistently well in nearly all conditions. The KLT (subspace) and logMMSE algorithms performed equally well, followed by the WT and MB algorithms. In babble noise (0 dB SNR), the KLT and Wiener-as algorithms performed the best among all algorithms.

3. The algorithms that were found in our previous study [5] to perform the best in terms of overall quality, were not the same algorithms that performed the best in terms of speech intelligibility. The KLT (subspace) algorithm was found in [5] to perform the worst in terms of overall quality, but performed well in the present study in terms of preserving speech intelligibility. In fact, in babble noise (0 dB SNR), the KLT algorithm performed significantly better than the logMMSE algorithm, which was found in [5] to be among the algorithms with the highest overall speech quality.

4. The Wiener-as algorithm performed the best in terms of preserving speech intelligibility (in one case, it improved intelligibility). We believe that this is due to the fact that it applies the least amount of attenuation to the noisy signal, thereby introducing negligible speech distortion. This is done, however, at the expense of introducing noise distortion (residual noise). At the other extreme, the pKLT approach reduces significantly the noise distortion but introduces a great deal of speech distortion, which in turn impairs speech intelligibility. In between the two extremes of speech/noise distortion lie the KLT and logMMSE algorithms.

5. The performance of speech enhancement algorithms, in terms of speech intelligibility, seems to be dependent

on the temporal/spectral characteristics of the noise, and this dependence is more evident in the low-SNR conditions (0 dB in our case). In the 0-dB train condition, for instance, none of the evaluated speech enhancement algorithms preserved speech intelligibility. The same algorithms, however, did preserve speech intelligibility in other noise conditions (same SNR).

Finally, it is important to point out that the disappointing conclusion drawn from this study that enhancement algorithms do not improve speech intelligibility is only applicable to normal-hearing listeners and not necessarily to hearing-impaired listeners wearing hearings aids [4] or cochlear implants [18]. In a different study [18], we showed that the KLT algorithm can significantly improve speech intelligibility in cochlear implant users. Further research is therefore needed to investigate the performance of speech enhancement algorithms in hearing-impaired listeners.

6. ACKNOWLEDGEMENTS

The authors would like to thank Jessica Dagley for her help in collecting the data. The authors also would like to thank Dr. Israel Cohen for providing the MATLAB implementation of the log-MMSE method with speech presence uncertainty.

7. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.

[2] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 26, pp. 471–472, 1978.

[3] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 479–514, Nov. 1997.

[4] K. H. Arehart, J. H.L. Hansen, S. Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," *Speech Communication*, pp. 575–592, 2003.

[5] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, pp. 153–156.

[6] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, pp. 225–246, 1969.

[7] ITU-T P.862, *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.

[8] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.," in *ISCA ITRW ASR2000*, Sept. 2000, Paris, France.

[9] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Proc.*, pp. 334–341, July 2003.

[10] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Proc.*, vol. 11, pp. 700–708, 2003.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[12] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, pp. 12–15, Jan. 2002.

[13] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Proc.*, pp. 799–807, 2001.

[14] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002.

[15] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Proc.*, pp. 59–67, Jan. 2004.

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

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

[18] P. C. Loizou, A. Lobo, and Y. Hu, "Subspace algorithms for noise reduction in cochlear implants," *The Journal of the Acoustical Society of America*, pp. 2791–2793, 2005.