A COMPARISON OF NON-INTRUSIVE SNR ESTIMATION ALGORITHMS AND THE USE OF MAPPING FUNCTIONS

James Eaton, Mike Brookes and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College, London, UK

{j.eaton11, mike.brookes, p.naylor}@imperial.ac.uk

ABSTRACT

We present a comparative evaluation of six methods for non-intrusive Signal-to-Noise Ratio (SNR) estimation for narrow-band speech in noise. We demonstrate that the performance of all methods can be improved by applying a non-linear mapping function to their estimates of SNR. We have employed phrases built from the TIMIT speech corpus and noises from a broad range of sources including ITU-T P.501, NOISEX-92, and Soundjay. We compare the accuracy of the methods in estimating the SNR of both stationary and non-stationary noise and we conclude that with the mapping function, the best current methods can estimate the SNR to within approximately 3.5 dB for SNRs from -5 dB to 35 dB.

Index Terms: speech enhancement, SNR, noise estimation

1. INTRODUCTION

Noise estimation is a fundamental building block in robust speech detection and speech enhancement algorithms. It is employed in a wide range of hardware applications including automotive hands-free kits, hearing aids, teleconferencing terminals, set-top boxes, home automation systems, mobile telephones, smartphones, and personal digital assistants, and in industries such as law enforcement. In this paper we consider the non-intrusive case in which the Signal-to-Noise Ratio (SNR) of noisy speech is estimated from the noisy speech alone without prior knowledge of the clean speech signal or the noise separately. For non-intrusive SNR estimation to be useful for a range of important applications in speech processing, audio and acoustic signal processing such as [1, 2], it must be sufficiently accurate over a range of noise levels, noise types and a wide range of speech characteristics in the utterances. We specify SNR in this context to mean the ratio of the power in the active speech to the power in the noise over the entire duration of the utterance under Intermediate Reference System (IRS) filtering [3], or under IRS filtering with A-weighting. The IRS filtering is used to simulate narrow-band telephony channel bandwidth. The SNR may vary with time depending on both the signal level and the noise level, and will also depend on any filtering that is applied such as A-weighting. Let the noisy speech signal be expressed as

$$y(n) = x(n) + v(n) \tag{1}$$

where $x(n)$ is the clean speech signal and $v(n)$ is uncorrelated random noise. Let $x_a(n)$ be the active speech [4] without pauses such that $y_a(n) = x_a(n) + v(n)$. We therefore aim to estimate the *a priori* SNR, ξ as

$$\xi = \frac{E\{{x_a}^2(n)\}}{E\{v^2(n)\}} \tag{2}$$

where E is the expectation operator. Estimating SNR is difficult when the noise level is very high relative to the speech or vice versa, and we have therefore included levels between -5 and 35 dB, and both stationary and non-stationary noise sources in our tests.

Approaches to SNR estimation in the literature can be considered in two classes: The first class uses a Voice Activity Detector (VAD) [5] to identify when speech is present in the signal, estimating the power of the speech by determining where the signal exceeds a predefined threshold, and then sampling the noise during the pauses to estimate the noise level. The ratio of the estimated mean speech power to the estimated mean noise power is the estimated SNR. The other class of methods track the minimum of the magnitude spectrum over a moving window which provide biased estimates of the noise level. Bias compensation has been employed in several automatic noise reduction algorithms including Martin [6] and latterly Hendriks [7]. A recent innovation proposed in [8] dispenses altogether with the bias compensation and safety net in [7] and uses Speech Presence Probability (SPP) with fixed priors to determine how much of the current noisy speech or previous noise estimate to include in the noise estimate for the current frame. In order to determine the non-intrusive SNR, we can estimate either the mean noise power (or noise Power Spectral Density (PSD)) given noisy speech, or else estimate the mean clean speech power given noisy speech. Assuming an estimate of the noise $\hat{v}^2(n)$ has been obtained, the estimated SNR, $\hat{\xi}$ can be computed as

$$\hat{\xi} = \max\left(\frac{E\{\hat{x}_a^2(n)\}}{E\{\hat{v}^2(n)\}}, \epsilon\right) = \max\left(\frac{E\{y_a^2(n)\}}{E\{\hat{v}^2(n)\}} - 1, \epsilon\right). \tag{3}$$

or when only the noisy speech power and an estimate of the clean speech power is known:

$$\hat{\xi} = \frac{1}{\left(\max(\frac{E\{y_a^2(n)\}}{E\{\hat{x}_a^2(n)\}} - 1, \epsilon)\right)} \quad (4)$$

where $E\{\hat{x}_a^2(n)\}$ is the estimated active clean speech power, ϵ is a suitable minimum value chosen to prevent $\hat{\xi}$ becoming negative, and $E\{\hat{v}^2(n)\}$ is computed from the average power in each frame of the PSD and applying Parseval's theorem. We can therefore derive $\hat{\xi}$ from the active noisy speech power $E\{y_a^2(n)\}$ and the noise PSD or the active speech level $E\{\hat{x}_a^2(n)\}$ provided by an estimator.

1.1. Benchmark testing

In order to make a broad meaningful comparison, we have chosen state-of-the-art algorithms using a range of different techniques, and some established algorithms for baseline comparison. The algorithms we have chosen are:

- ACT, an implementation [9] of ITU-T P.56 [4] that estimates speech activity level and duty cycle. The algorithm returns the speech activity level and activity factor. The SNR is obtained from Eq. (4) using the estimated speech power obtained by multiplying the active speech power with the activity factor, and the noisy speech power. This algorithm is not intended for use with low SNRs since noise occurring during the noisy speech is included in the active speech power estimate.

- STN, the National Institute Of Standards And Technology's stnr SNR estimation algorithm [10]. The algorithm returns the estimated SNR directly.

- KSM, an implementation of Kim and Stern's SNR estimation algorithm [11] in Matlab which interpolates the values in the look up table providing an improvement in estimation accuracy. The algorithm returns the estimated SNR directly.

- EST, an implementation of [6], with Table 3 replaced by the updated Table 5 from [12]. A slight improvement was reported by Mauler and Martin [13] but this is not included [9]. The SNR is obtained from Eq. (3) using the estimated noise power and the noisy speech power.

- ESG, an implementation [9] of Gerkmann's noise PSD tracker [8] providing a reduction in computational complexity. The SNR is obtained from Eq. (3) using the estimated noise power and the noisy speech power.

- HEN, R. Hendrik's Minimum Mean Squared Error (MMSE)-based noise PSD tracker with low complexity [7] Matlab code. The SNR is obtained from Eq. (3) using the estimated noise power and the noisy speech power.

Previous comparative evaluations include Cohen [14] using twenty utterances and noise between -5 and $10\,$dB; Beritelli [15] considers vowel sounds only; Vondrasek [16] using similar test data. Hendriks [7] compares five algorithms in non-stationary noise in the range 0 to $15\,$dB; Ren [17] uses a sample size of ten utterances. Taghia [18] compares eight algorithms in a range of noise types using two speech signals of two-minute duration.

The contribution of this paper is to extend the evaluation to a wider set of noises spanning stationary and non-stationary examples, and to include speech pauses in the test utterances such as may occur in natural speech but are not always included in previously used test databases. Furthermore, we have included in our study a wider range of SNR levels than has previously been investigated. This is consistent with one of our motivations which is towards the application of surveillance in law enforcement for which negative SNRs are common. The six algorithms investigated span a broad range of methods evaluated in terms of their medians, and inter-quartile ranges, and in addition we have determined the accuracy based on 95% confidence intervals.

1.2. Mapping of instrumental measures to SNR

To increase accuracy, we propose the use of a novel mapping function between the algorithms' outputs and the true SNR and we have determined this function for each algorithm. This mapping was trained on the training corpus of the TIMIT [19] database, and stationary and non-stationary training noises were obtained from the ITU-T P.501 [20] monaural noise sequences. Coefficients for a fifth order polynomial fit in a least squares sense were computed over the range $-5\,$dB to $35\,$dB dB in $5\,$dB steps from the estimated SNR from each algorithm to minimise the maximum dB error against the true SNR at these points.

2. TEST METHODOLOGY

Phrases from the TIMIT [19] training and test databases simulating realistic clean speech were assembled. Composite utterances comprising three concatenated TIMIT files separated by one second pauses were constructed totalling 448 speech files for the test dataset and 1,152 speech files for the training dataset. TIMIT sentences SA1 and SA2 were excluded since these are repeated for each speaker. Stationary test dataset noises were obtained from the NOISEX-92 noise database [21]. Non-stationary test dataset noises were obtained from Soundjay [22] and modulated white noise and Simultaneous Switching Noise (SSN) provided by R. Hendriks as used in [7].

To provide narrowband telephony channel conditions, the IRS [3] weighting was applied to the noisy speech signals. Clean speech phrases and noises truncated to the length of the speech were resampled to $f_s = 8\,$kHz and mixed in SNRs calculated using the mean speech power derived from ITU-T P.56 [4] and the mean noise power. Actual SNRs ranging from -5 to $35\,$dB in $5\,$dB increments were used for the test dataset, and with -6 to $36\,$dB in $5\,$dB increments for the training dataset. The A-weighted SNR was also calculated and is

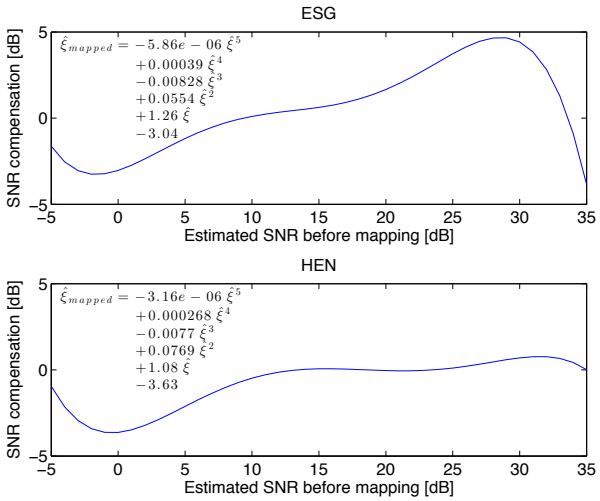

Fig. 1. *Trained mapping functions for the ESG and HEN algorithms.*

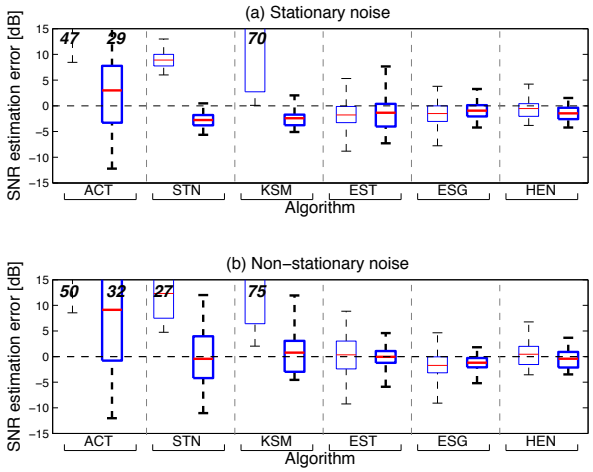

Fig. 2. *Box plots [23] for SNR estimation error over the range* -5 *to* 35 dB *by algorithm before and after mapping. Numbers in bold italics indicate extents of whiskers not shown.*

shown in the overall comparison. The coefficients from the polynomial fit were then evaluated using the output from each algorithm under test. Fifth order was found to be a useful improvement over lower orders of polynomial, and better performing than cubic or spline interpolation in our tests, whilst higher orders than five did not increase accuracy.

Tests were performed covering the permutations of algorithm, noise and actual SNR for each dataset. 1,800 noisy speech samples were tested with each algorithm for each of the test and training databases and for SNR and A-weighted SNR comprising 14,400 tests in total. We also computed the 95% confidence intervals for the SNR estimation errors in dB bounded by the 2.5th and 97.5th centiles.

The estimated Real-Time Factor (RTF) defined as processing time divided by the duration of the speech signal for each algorithm was determined by measuring the elapsed processing time using the Matlab *cputime* function for each call on a 2.3 GHz Intel i5 Core processor with 4 GB 1.333 GHz DDR3 memory, and calculating the mean time per algorithm divided by the mean speech file duration.

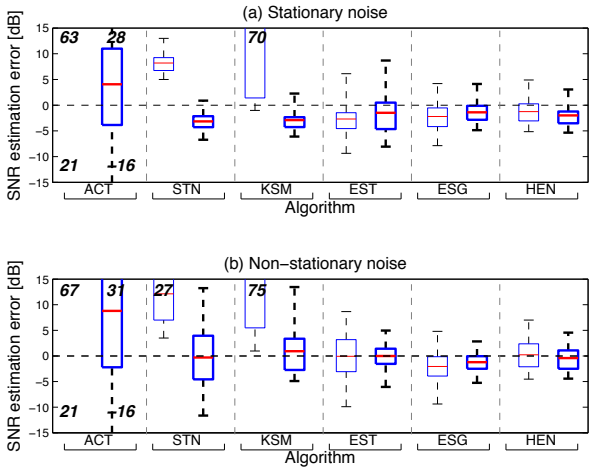

Fig. 3. *Box plots for SNR estimation error over the range* -5 *to* 35 dB *by algorithm with A-weighting before and after mapping. Numbers in bold italics indicate extents of whiskers not shown.*

3. RESULTS AND DISCUSSION

The characteristics and coefficients for $\hat{\xi}_{mapped}$ of the mapping functions based on training data for the two best performing algorithms in stationary noise HEN and ESG are shown in Fig. 1. This illustrates the similar performance of HEN and ESG at SNRs of 15 dB and below, and the larger estimation errors at higher SNRs produced by ESG. ESG is most accurate at 15 dB, the optimal fixed prior $10\log_{10}(\xi_{\mathcal{H}_1})$ found in [8]. HEN is least accurate at 0 dB, the point at which the bias function, B in [7] increases. The mapping functions here have been trained on SNRs between -5 and 35 dB, therefore the estimation performance is uncertain outside of this range. Estimating very high or low SNRs is difficult since either the speech energy is very much greater than the noise or vice versa and therefore where algorithms typically fail, and training on a wider range will therefore alter the mapping within the range tested here.

Figs. 2 and 3 show the results by algorithm for SNR and A-weighted SNR respectively before and after mapping. Results after mapping are plotted in bold. The box plots [23] show the medians, 25th and 75th centiles, and the whiskers extend to the 2.5th centile and 97.5th centiles between which 95% of the data lies. From Fig. 2 it can be seen that the map-

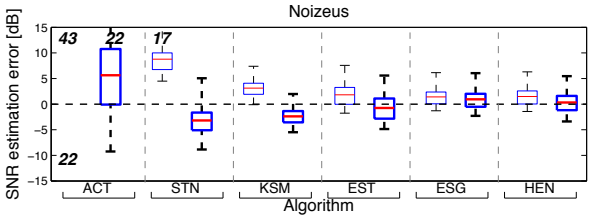

Fig. 4. *Box plots for SNR estimation error using the NOIZEUS [24] corpus over the range* 0 *to* 15 dB *by algorithm before and after mapping. Numbers in bold italics indicate extents of whiskers not shown.*

ping technique substantially improves the accuracy and reduces the bias of all algorithms for both noise types. This is mainly due to large variances at low SNRs. The ACT method estimates have a large error as the algorithm estimates speech power including any noise present during active speech. STN performs well in stationary noise with 95% of estimations falling between $+1.8$ and -3.4 dB. With non-stationary noise however it performs poorly: the negative excursion at -5 dB not shown is approximately -25 dB. EST gives better performance on non-stationary sources whilst performing slightly less well on stationary data. HEN provides the best performance overall with a spread of data of approximately 7.3 dB over the 95% of results. However, ESG provides similar performance except at high SNRs but with much reduced computational complexity as will be discussed later in this section. Fig. 3 shows that the results for A-weighted SNR estimation are very similar to the unweighted results for all algorithms.

For further verification of the results, the algorithms were tested on the NOIZEUS [24] corpus but using the TIMIT-trained mapping functions. NOIZEUS comprises speech in noise at SNRs of 0, 5, 10, and 15 dB. The noises used were: airport; babble; car; exhibition; restaurant; station; street; and train; a mixture of both stationary and non-stationary noise. The 95% confidence intervals for the NOIZEUS tests are shown in Fig. 4. The SNR estimation performance ranking of the algorithms was similar to the TIMIT [19] stationary results, and as expected due to the mixture of stationary and non-stationary noise, algorithms which give large errors below 0 dB and above 15 dB in non-stationary noise show good results in this test.

Fig. 5 and Fig. 6 show the SNR estimation errors after mapping calculated over the full range of test SNRs for the three best performing algorithms in stationary noise, STN, ESG, and HEN.

Table 1 shows the estimated RTFs for each of the algorithms using the method discussed in Sec. 2. The ESG method has a similar RTF to STN but superior SNR estimation performance in non-stationary noise. The HEN method has the highest RTF due to the safety net and bias compensation com-

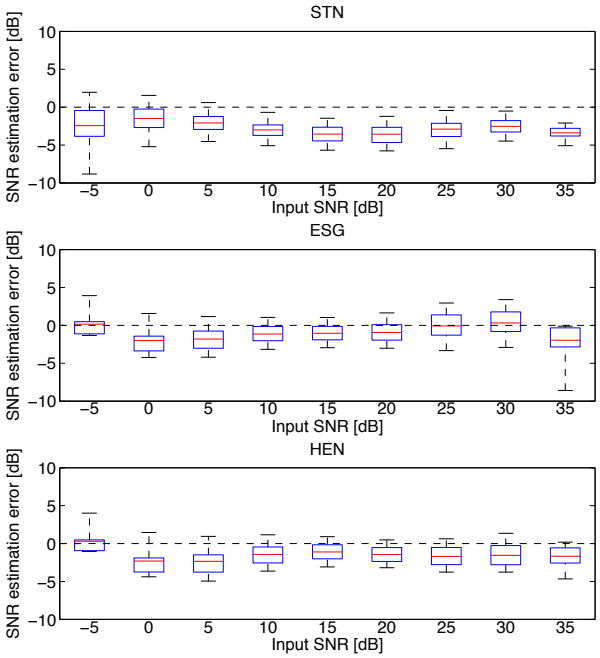

Fig. 5. *Box plots of estimation error by input SNR in stationary noise for the STN, ESG, and HEN algorithms after mapping*

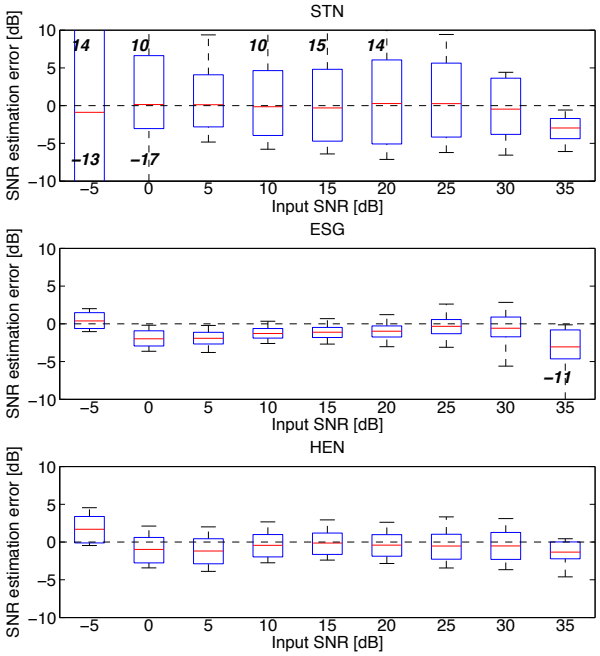

Fig. 6. *Box plots of estimation error by input SNR in non-stationary noise for the STN, ESG, and HEN algorithms after mapping. Numbers in bold italics indicate extents of whiskers not shown.*

putations in the algorithm. In most circumstances therefore

ESG may therefore be preferable to HEN except where good performance at high SNRs is required. All implementations

Table 1. *Real-Time Factor by algorithm.*

ACT	STN	KSM	EST	ESG	HEN
0.027	0.0045	0.0080	0.012	0.0053	0.35

were in Matlab except for STN which was a 64-bit compiled executable.

4. CONCLUSIONS

We have presented a quantitative comparison of non-intrusive SNR estimation algorithms and introduced a trained mapping function. Tests were conducted over a wide range of input SNRs and noise types including both stationary and non-stationary noise. Our results show that for narrowband speech in noise over the range of input SNRs from -5 to $35\,\mathrm{dB}$ the best performing algorithm is HEN [7], which has an ability to estimate noise to within approximately $\pm 3.5\,\mathrm{dB}$. The ESG method however with only slightly reduced accuracy to HEN has significantly lower computational complexity, and can outperform HEN in some situations.

We have demonstrated that the use of a non-linear mapping function can reduce both the bias and variance of the measurement error and that, with the inclusion of this mapping function, 95% confidence intervals are reduced by around $5\,\mathrm{dB}$ for the ESG method and $3\,\mathrm{dB}$ for the HEN method.

5. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[3] ITU-T, *Transmission Standards Recommendation P.48 Specification for an intermediate reference system*, International Telecommunications Union (ITU-T) Recommendation P.48, Nov. 1988.

[4] ——, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[5] M. Grimm and K. Kroschel, Eds., *Robust Speech Recognition and Understanding*. I-Tech, Vienna, Austria, 2007.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[7] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 4266–4269.

[8] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[9] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2013.

[10] NIST, "NIST speech tools," http://www.nist.gov/speech/tools, 2006.

[11] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech Conf.*, Sep. 2008, pp. 2598–2601.

[12] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, Jun. 2006.

[13] D. Mauler and R. Martin, "Noise power spectral density estimation on highly correlated data," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Sep. 2006.

[14] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 9, pp. 725–728, Sep. 2004.

[15] F. Beritelli, S. Casale, R. Grasso, and A. Spadaccini, "Performance evaluation of SNR estimation methods in forensic speaker recognition," in *Emerging Security Information Systems and Technologies (SECURWARE), 2010 Fourth International Conference on*, Jul. 2010, pp. 88–92.

[16] M. Vondrasek and P. Pollak, "Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency," *Radio Engineering*, vol. 14, pp. 6–11, Jan. 2005.

[17] Y. Ren and M. Johnson, "An improved SNR estimator for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 4901–4904.

[18] J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4640 –4643.

[19] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[20] ITU-T, *Test signals for use in telephonometry*, International Telecommunications Union (ITU-T) Recommendation P.501, Aug. 1996.

[21] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[22] Soundjay.com, "Sound effects website," http://www.soundjay.com/.

[23] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of boxplots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[24] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 153–156.