Evaluation of Instrumental Measures for the Prediction of Musical Noise in Enhanced Noisy Speech

M. R. J. Gerrits[1], R.C. Hendriks[1], N. D. Gaubitch[1], J. Jensen[2,3], and M. S. Pedersen[2]

[1] Delft University of Technology, 2628 CD, Delft, The Netherlands
[2] Oticon A/S, 2765 Smørum, Denmark
[3] Aalborg University, 9100 Aalborg, Denmark

Abstract. Instrumental measures have been primarily used in order to predict speech quality or speech intelligibility. The aim of the present work is to evaluate the performance of a broad range of established instrumental measures in terms of their ability to predict the amount of musical noise present in enhanced noisy speech signals. The considered instrumental measures were evaluated using musical noise quantity scores obtained from a specially designed listening-test which was performed by normal-hearing listeners. Of all considered instrumental measures, a mean squared distortion measure, the PESQ measure, and the STOI measure yielded the highest correlations. These results confirm the ability of instrumental measures to predict the amount of musical noise, but further evaluation shows limitations to their applicability as musical noise predictors. However, subsequently the results suggest that simultaneous optimization for the amount of musical noise and speech quality or speech intelligibly is not possible using a single instrumental measure.

1 Introduction

The interest in the prediction of speech quality and speech intelligibility has led to the development of multiple instrumental speech-quality and speech-intelligibility measures. This development is supported by the demand for replacement of time consuming and expensive listening tests. However, as instrumental measures are not perfect, both instrumental measures as well as listening tests, are essential to judge the performance of speech enhancement algorithms. Over the years, many noise suppression algorithms have been introduced attempting to find optimal noise suppression without introducing loss of speech-quality or intelligibility under a diversity of distortion measures, see e.g., [1,2]. The performance of such speech enhancement algorithms is in general a trade-off between reduction of the noise and the extent in which the algorithm distorts the speech signal. Finding the right tradeoff which is crucial in the field of hearing aids, where the sound to be presented to the hearing impaired listener should be as natural and with as high quality as possible.

However, removing all the noise in a noisy speech signal without introducing loss of speech quality requires that the noise is known exactly. This is unrealistic for any practical application. Therefore, many noise reduction methods employ the statistics of the speech and noise processes, e.g. the speech and noise power spectral densities. As a result of working with statistical descriptors of the signal, instead of the actual realizations, artifacts are introduced during the noise reduction process. Among these is the highly annoying residual noise known as musical noise [3], which can remain after processing, and decrease the quality and speech intelligibility of the enhanced speech signal. Although it is possible to hide or reduce the musical noise to some extent by adjusting certain parametric settings of the noise reduction algorithm [3,4], this is not straightforward without a clear instrumental measure.

While some instrumental measures have shown to predict the quality of enhanced noisy speech with high correlation (see [5], [6], [7]), only little is known about the ability of instrumental measures to predict the amount of musical noise, even though this is an important aspect of speech quality. The aim of this contribution is to evaluate the ability of an instrumental measures to predict the amount of musical noise in the enhanced signal. Given that there are measures which are able to predict the amount of musical noise, these could be used to draw additional conclusions about the performance of noise reduction methods and be of use in their development.

In 2008, Uemura et. al [8] proposed a novel measure based on the kurtosis of the power spectral density (PSD) of the enhanced noisy speech signal which can be used specifically for predicting the amount of musical noise in the enhanced speech signal. Besides this measure, we consider various other established instrumental measures in terms of their ability to predict the results of a specially designed listening experiment. The results of the subjective listening-test characterize the quantity of musical noise in several enhanced noisy speech signals generated by applying the spectral subtraction noise reduction method (SS), and varying the amount of noise reduction. Evaluation of the prediction performance of the different instrumental measures is based on Pearson's correlation coefficient and the Kendall rank correlation coefficient [9].

2 On spectral subtraction and musical noise production

We consider a noisy speech model that can be described as a clean speech signal $s(n)$ that is degraded by an additive uncorrelated noise signal $v(n)$, i.e., $y(n) = s(n) + v(n)$, where $y(n)$ represents the observed noisy speech with time-sample index n. It is assumed that both $s(n)$ and $v(n)$ are wide-sense stationary for short time segments, and a DFT based analysis-modification-synthesis procedure is applied, with the aim of reducing noise. After applying a short-time Fourier transform (STFT) to the noisy signal, the noisy observation can be described by,

$$Y(i,k) = S(i,k) + V(i,k), \tag{1}$$

where upper case letters describe the signal in frequency domain. The frame-index and transform coefficients are described by i and k respectively. The spectral subtraction noise reduction method [3, 10] can be used to modify the frequency spectrum of the noisy observation. By subtracting a noise PSD estimate, $\sigma_V^2(i,k)$, from a noisy observation, $|Y(i,k)|^2$, a clean speech periodogram, $|\widehat{S}(i,k)|^2$, is estimated. The SS noise reduction procedure is described by,

$$|\widehat{S}(i,k)|^2 = \begin{cases} H(i,k)|Y(i,k)|^2 & \text{for } H(i,k)|Y(i,k)|^2 > \alpha_{ss}\sigma_V^2(i,k) \\ \alpha_{ss}\sigma_V^2(i,k) & \text{otherwise} \end{cases} \tag{2}$$

where $H(i,k)$ is given by,

$$H(i,k) = 1 - \frac{\beta_{ss}\sigma_V^2(i,k)}{|Y(i,k)|^2}. \tag{3}$$

β_{ss} is called the over-subtraction factor and represents a scalar weight on $\sigma_V^2(i,k)$, α_{ss} denotes a flooring parameter which introduces a minimum value for gain, $H(i,k)$. After modification the signal is reconstructed using an inverse STFT.

However, even when the noise PSD is perfectly known, it is impossible to remove all the noise as this requires all noise realizations to be known. These variations of the power of these noise realizations around its PSD will introduce estimation errors in $H(i,k)$. These estimation errors result in variations in $H(i,k)$ with musical noise as a consequence. Consider a noise-only segment, $y(n) = v(n)$, where $|V(i,k)|^2 > \beta_{ss}\sigma_V^2(i,k)$ thus $H(i,k) > 0$, this means that part of the noise power will remain in the enhanced periodogram. If this power residue is isolated in both time and frequency, i.e. a narrow-band spectral peak, it will introduce a short tonal artifact after reconstruction of the signal. Whenever isolated spectral peaks occur in multiple time-frames across different frequency bands, the residual noise is described as musical noise. Generally, altering β_{ss} and α_{ss} provides a trade-off between the amount of musical noise, broadband noise reduction and speech distortion.

3 Instrumental measures

A broad range of established instrumental measures is used with the objective to evaluate their ability to predict the amount of musical noise in enhanced noisy speech. Included are the often used Perceptual Evaluation of Speech Quality measure, the log-likelihood ratio, the segmental SNR, the frequency weighted segmental SNR, and the normalized frequency weighted segmental SNR. But also several spectral distance measures are taken into account, i.e. the weighted spectral slope, the cepstral distance, the log-spectral distance, two mean squared error measures, the Itakura-Saito distance, a COSH distance measure based on the IS measure, and two composite measure for noise distortion, which is based on a combination of seven objective SQ measures. Additionally we apply one musical noise measure, based on the kurtosis-ratio between the PSD of the enhanced and the noisy speech signal, and one speech-intelligibility measure, the short-time objective intelligibility measure. Table (1) provides a summary of all instrumental measures with their corresponding abbreviations. For the PESQ, LLR, fwSEG, fwSEGn, IS1, CEP, WSS and COMP measure we apply the implementations as provided by Loizou in [2], whereas for IS2, segSNR and the COSH metric we apply the implementation as provided in the VOICEBOX toolbox [11]. The IS1 and IS2 measures are the same, however IS1 is an LPC based implementation, where IS2 is calculated using the PSD of the clean and enhanced speech. Both these IS measures and the COSH measure have been limited to 100. The two different Mean Squared Distance methods, d_{MSD1} and d_{MSD2}, are implemented based on respectively the complex DFT coefficients ($X = S$), and the periodogram ($X = |S|^2$) of the estimated clean and noisy speech signal.

	Instrumental Measures
PESQ	Perceptual Evaluation of Speech Quality [12]
LLR	Log-Likelyhood ratio [13]
segSNR	Segmental SNR [13]
fwSNR	Frequency weighted segmental SNR [13]
fwSNRn	normalized Frequency weighted segmental SNR [5]
LSD	Log-Spectral Distance [14]
MSD1	Mean Squared Distance
MSD2	Mean Squared Distance (PSD)
IS	Itakura-Saito Distance [13]
COSH	Symmetric version of IS
CEP	Cepstral Distance [13]
WSS	Weighted Spectral Slope [13]
STOI	Short-Time Objective Intelligibility Measure [14]
KURT	Kurtosis ratio [8]
COMPovl	Composite Noise Distortion Measure, overall quality [5]
COMPbn	Composite Noise Distortion Measure, background distortion [5]

Table 1. The various different instrumental measures used for evaluation

$$d_{MSD}(X) = \frac{1}{M}\sum_k \sqrt{\frac{1}{L}\sum_i |X(i,k) - \widehat{X}(i,k)|^2} \tag{4}$$

where M describes the total number of frequency coefficients, and L denotes the total number of time-frames. The 'kurtosis ratio' measure is described by the ratio between the kurtosis of the PSD of the noisy signal, $kurt_{noisy}$, and the kurtosis of the PSD of the enhanced signal, $kurt_{enhanced}$, i.e.,

$$d_{kurtosis-ratio} = \frac{\frac{1}{M}\sum_k kurt_{enhanced}(k)}{\frac{1}{M}\sum_k kurt_{noisy}(k)}, \tag{5}$$

where *kurt* is defined by,

$$\mathit{kurt(k)} = \frac{\frac{1}{L}\sum_i X(i,k)^4}{(\frac{1}{L}\sum_i X(i,k)^2)^2} \tag{6}$$

Dependent on whether $kurt_{noisy}$ or $kurt_{enhanced}$ is computed, X represents either the realizations of the noisy signal, $|Y(k)|^2$, or the realizations of the enhanced speech signal, $|\widehat{S}(k)|$. In [8] it was stated that a high kurtosis-ratio corresponds to a high amount of musical noise, i.e. the distribution of the isolated residual noise coefficients correspond to a high kurtosis values, which indicates a distribution with a sharp peak. However, this measure contains some practical limitations, e.g. the proposed measure is solely applied on noise-only segments, [8]. This is solved using a binary mask based on a SNR threshold of 0 dB. Also note that calculating the frequency average in Eq. (4,6), does only make sense for distributions which are spectrally flat.

4 Listening experiments

The data-set, Q, used for this experiment includes a set of 20 different clean speech estimates, $\widehat{s}(n)$. These clean speech estimates are generated by applying the SS noise reduction method, Eq. (2), to a noisy speech signal, and varying the subtraction-factor β_{ss}. Hence, these enhanced signals contain different amounts of musical noise. As masking of residual background noise is undesired for this specific experiment, the spectral floor, α_{ss}, is set to 0.

The clean speech, $s(n)$, used for generating the stimuli, consisted of a concatenation of two sentences which originated from the TIMIT [15] database sampled at 16 kHz. The two different sentences where respectively read by a female and a male, with a combined total signal length of approximately 7 seconds.

A noisy speech signal, $y(n)$, is produced by degrading this clean speech signal with additive white Gaussian noise, $v(n)$, and the signal-to-noise ratio is set to 5 dB. The noise PSD is estimated over 10 seconds of noise-only signal. All noise reduction is performed in the power spectral domain, using time frames of 32 ms with 50% overlap and a 512 point FFT is used to transform each time-frame into the DFT domain, and a square-root Hann window is used as analysis and synthesis window. As stated before, altering the over-subtraction factor, β_{ss}, allows changing the amount of musical noise. The over-subtraction factor is varied in steps of 0.5 from $\beta_{ss} = 0$, i.e. no noise reduction, to $\beta_{ss} = 9.5$ where the estimated clean speech consists of highly distorted speech and little to no musical noise due to the large over-subtraction. Such a large value of β_{ss} is rarely in practice due to the amount of speech distortion that it introduces. However, it is an important for a musical noise predictor to be able to identify situations in which little or no musical noise remains.

The listening test is designed to characterize the amount of musical noise in an enhanced noisy speech signal. Listeners were asked to specifically grade the amount of musical noise, by answering the following two, equivalent, questions: 'How much musical noise is present in the stimuli? That is, How musical is the background noise perceived?'. However, comparison of the 20 enhanced signals at once is too tiring for the listener, therefore, the data-set, Q, is split into two different subsets of each 10 enhanced speech signals, Q_{set1} and Q_{set2}, respectively containing the enhanced signals, $\widehat{s}(n)$ generated with over-subtraction factors $\beta_{ss} \in \{0, 1, 2, ..., 9\}$, and $\beta_{ss} \in \{0.5, 1.5, 2.5, ..., 9.5\}$.

All subjects preformed a training round, which consists of subset Q_{set2} and two test rounds, where Q_{set1} and Q_{set2} are successively presented to the listener. The signals in each subset are presented simultaneously, compared among each other, and graded individually by assigning a single number in the range 1 to 5 with decimal steps of 0.1, providing a numerical indication of the amount of musical noise. That is, the listeners rate the amount of narrow-band residual noise present in the background of the presented stimuli by comparing the 10 presented signals. Grade 1 is assigned to signals with no musical noise e.g. broadband noise or no background distortion, while grade 5 is assigned to signals with extreme amounts of musical noise. The ratings are depicted in Table 2.

score	**The amount of musical noise**
5	Extreme
4	A lot
3	Medium
2	A little
1	Broadband or no musical noise

Table 2. Grading scale for the amount of musical noise

Prior to the actual listening test, 3 signals are presented to the listener to make the subject accustomed to the signals presented in the listening test. Two of these signals contain enhanced noisy speech with some degree of musical noise while the other signal contains the unprocessed noisy speech. Subjects are asked to explain the task at hand, to ensure complete understanding of the test. The listening test was performed in total by 20 subjects inside a sound proof listening room. The group of subjects consists 90% male, and 10% female listeners, within the age-range of 23 up to 51 years old. The average age of a listener was 30 years old, and the listeners were not hearing impaired to the best of their knowledge. Finally a mean opinion musical noise quantity score is calculated by averaging over the obtained scores. The average scores of the listening tests are depicted Figure 1.

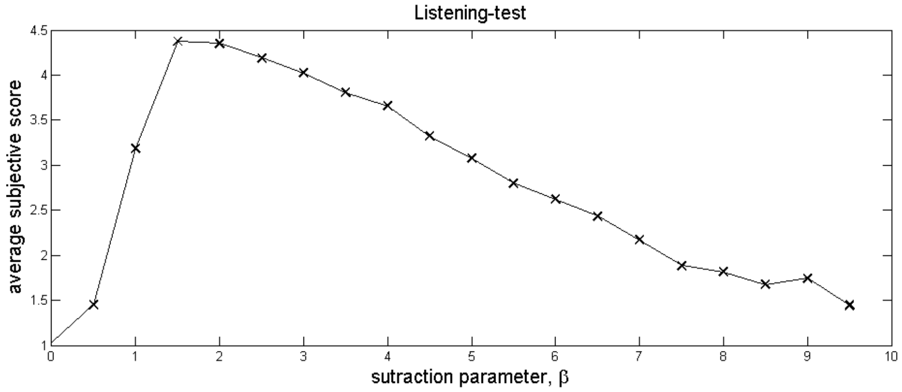

Fig. 1. Listening-test results for quantifying musical noise, i.e. The mean-opinion musical noise quantity score.

5 Evaluation

All instrumental measures (Table 1) are evaluated using data-set Q, after which we apply two different performance metrics to measure the performance in terms of their ability to predict the amount of musical noise. Performance measures are commonly used for evaluating correlations between instrumental and listening test data, see [2]. First, the Pearson correlation coefficient, ρ, is included,

$$\rho = \frac{\sum_l (S_l - \overline{S})(D_l - \overline{D})}{\sqrt{\sum_l (S_l - \overline{S})^2 \sum_l (D_l - \overline{D})^2}}. \tag{7}$$

Here, S and D describe, respectively, the mean opinion musical noise quantity scores (Figure 1) and the instrumental scores. $\overline{S}$ and $\overline{D}$ define the means of the sets S and D, and l denotes the over-subtraction index. Secondly we imply Kendall's tau rank correlation coefficient,

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}. \tag{8}$$

Here, N_c describes the concordant pairs, while N_d describes the discordant pairs [9]. N denotes the total number of enhanced speech signals applied for the subjective test, i.e. $N = 20$. This latter measure is used to support the conclusions made based on the correlation coefficient ρ.

To compare the performance of the different instrumental measures, the absolute value of both coefficients is applied, i.e. $|\rho|$ and $|\tau|$ must be in the range of $0 < |\rho| < 1$ or $0 < |\tau| < 1$. Here, a coefficient close to one corresponds to a high agreement between the instrumental measure and the subjective data, whereas a coefficient close to zero indicates the data to be independent. An overview of the performance of all instrumental measures is shown in Figure 2.

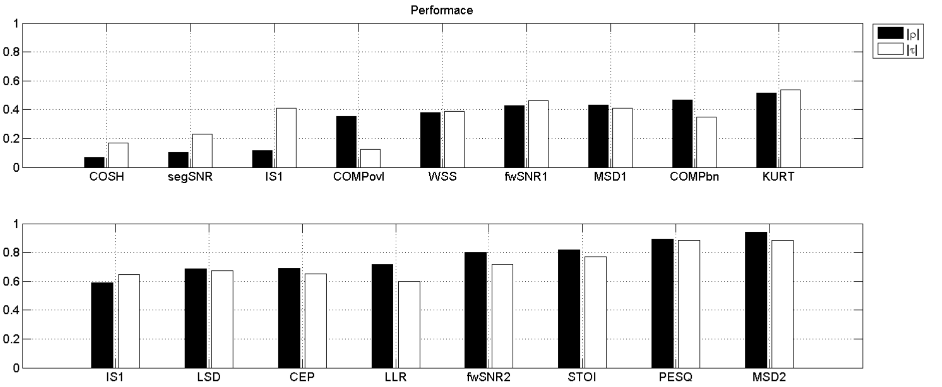

Fig. 2. Performance overview, the ability of instrumental measures to predict the amount of musical noise.

6 Results and Discussion

The results displayed in Figure 2 indicate that the STOI, PESQ and MSD2 measure provide a good prediction of the amount of musical noise present in the enhanced signal, there a strong correlation exists between the output of these instrumental measures and the listening test scores. Therefore, we will evaluate the performance of these three methods in more detail. The results of these three metrics are depicted in Figure 3, illustrated by the dashed line. To gain insight into how the gain function, $H(i, k)$, distorts the clean speech, we apply gain $H(i, k)$ to the the clean speech $|S(i, k)|^2$, and calculate the results for the three instrumental metrics. These results are portrayed in Figure 3, and illustrated by the dotted line. And immediately a rather important observation can be made when comparing the dashed and dotted curves for each of the three measures. Although the dashed curves of the PESQ and the STOI measure provide a strong correlation with the listening test results, we can not conclude that these two metrics are good measures for predicting musical noise in general. Note that the dotted curves represent processed clean speech, which does not contain any musical noise. However, for these signals the PESQ and the STOI measure produce high results, and therefore it is not possible to predict the amount of musical noise without prior knowledge on the presence of musical noise. This limits the applicability of the PESQ and the STOI measure as a musical noise predictor drastically. Considering the performance of MSD2, it can be seen that this measure is sensitive to scaling of the enhanced noisy speech signal, there scaled versions of the enhanced noisy speech signal will produce different outputs. However, the perception of the sounds will not change.

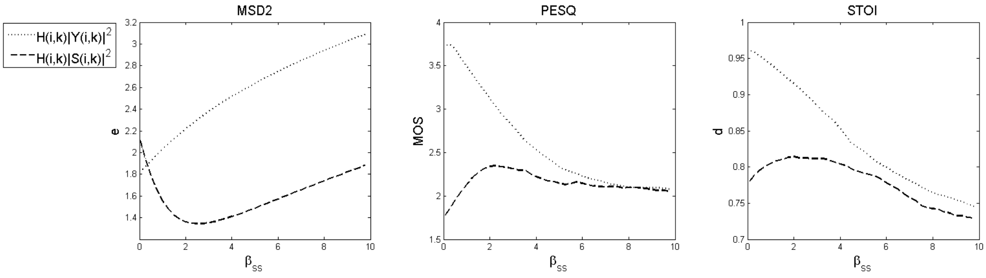

Fig. 3. Results of the three instrumental measures that show high performance for predicting musical noise

Consider the minimum of the MSD2 measure as a function of β_{ss}, here the average distance between $|S(i, k)|^2$ and $|\widehat{S}(i, k)|^2$ is small, i.e. the average distortion is small. But note, comparing Figure 1 and 3, that whenever the MSD2 at its minimum, the amount of musical noise is high. This indicates that the best clean speech estimate contains according to this instrumental measure the largest amount of musical noise. A similar conclusion can be drawn when evaluating the results of the PESQ measure, there the maximum mean opinion scores, i.e. the maximum speech quality scores, are predicted for β_{SS} values that lead to high amounts of musical noise. Likewise, in case of the STOI measure, the maximum speech intelligibility scores are predicted whenever the amount of musical noise is high according to Figure 1. This means that when optimizing the enhanced speech signal for intelligibility, quality or the distortion, the signal will contain a high amount of musical noise. Subsequently, optimizing for both the minimum amount of musical noise and the best speech quality simultaneously is not possible using the instrumental measures considered in this paper.

Another observation can be made, based on the results of the PESQ and STOI measure. From Figure 3 it can be seen that it is possible to apply high values of β_{ss} and largely reduce the amount musical noise, without loosing a lot of speech quality, and on account of introducing only a limited decrease in speech intelligibly.

Additionally we can conclude from Figure 2 that the background distortion measure, COMPbn, and the musical noise prediction method, KURT, do not correlate well with the listening test results. In case of the COMPbn measure this result is consistent with the observations made in [5]. However, in case of the KURT measure this is a more unexpected result, but can be explained as follows. First consider Equation (6). If β_{ss} is increased, then more and more spectral values will be set to zero, and the PDF of the enhanced speech will become increasingly peaky. As a consequence, the kurtosis of the PDF of the enhanced signal will increase until all spectral values are set to zero, for which the method becomes undefined. By comparing this with the results in Figure 1, it can observed that the increasing $d_{kurt-ratio}$ has a strong correlation for the range $0 < \beta_{ss} < 2$, which is consistent with the experiments in [8]. However, for the range $\beta_{ss} > 2$ the $d_{kurt-ratio}$ metric keeps increasing, where the listening test results show a decrease in the amount of musical noise. We can thus conclude that the KURT metric performs poorly for values of $\beta_{ss} > 2$.

Form Figure 2 we observe that a large group of instrumental methods show low correlation with the listening test results. This could be caused by the fact that both τ and ρ only provide a metric to illustrate a linear relation. In order to obtain a better correlation between the results of the instrumental measures and the listening test scores, a mapping could be used to account for nonlinear relations, see [14].

7 Conclusions

The results provided in this paper concluded that it is possible to predict the amount of musical noise in an enhanced noisy speech signal by either using a mean squared distance measure, MSD2, the PESQ measure or the STOI measure. There the results of these instrumental methods indicate strong correlations with the listening test results. However, this high correlation suggests that simultaneous optimization for the minimum amount of musical noise and maximum speech quality or speech intelligibly is not possible using a one of the considered instrumental measure. Surprisingly, the instrumental method, KURT (Table 1), which was specially designed for predicting of the amount of musical noise shows low correlation with the listening test results. On basis of the analysis that existing instrumental methods are limited in the prediction of musical noise, we can speculate on the need for a better metric specifically designed to measure the musical noise quantity.

References

1. R. C. Hendriks, T. Gerkmann, and J. Jensen. *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement, A Survey of the State-of-the-Art.* Morgan and Claypool Publishers, 2013.
2. P. Loizou. *Speech enhancement theory and practice.* CRC Press, 2007.
3. M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, volume 4, pages 208–211, 1979.
4. O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Processing*, 2(2):345–349, April 1994.
5. Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.
6. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time frequency weighted noisy speech. *IEEE ICASSP*, 19(7):2125–2136, 2011.

7. C. H. Taal, R. C. Hendriks, Heusdens, and J. Jensen. An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *J. Acoust. Soc. America*, 130(5):3013–3027, 2011.
8. Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo. Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation. pages 4433–4436, 2009.
9. D. J. Sheskin. *Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 3rd edition edition, 2004.
10. S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.
11. Voicebox: Speech processing toolbox for matlab. *http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, May 2013.
12. ITU-T P.862. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs. Technical report, 2000.
13. S. R. Quackenbush. Objective measures of speech quality. pages 44–57, 1995.
14. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE ICASSP*, pages 4214–4217, 2010.
15. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium, Philadelphia, 1993.