EVALUATION OF INSTRUMENTAL MEASURES FOR THE PREDICTION OF MUSICAL NOISE IN ENHANCED NOISY SPEECH

By

M.R.J. Gerrits

This thesis has been prepared under supervision by,

Dr. Ir. R.C. Hendriks (Delft University of Technology)
Dr. N.D. Gaubitch (Delft University of Technology)
Dr. M.S. Pedersen (Oticon A/S)
Prof. Dr. J. Jensen (Oticon A/S & Aalborg University)

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in the School of Engineering at
Delft University of Technology, 2014

Delft, The Netherlands

Abstract. To obtain the absolute truth about the performance of a noise reduction method one requires to perform a listening experiment. As listening experiments are often time consuming and expensive there exists a need to replace these experiments by instrumental measures. Consequently, research has provided various instrumental measures which can been 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 performance of the instrumental measures is evaluated using musical noise quantity scores obtained from a specially designed listening experiment which was performed by normal-hearing listeners. The investigated stimuli, which contain various amounts contain musical noise, are produced using the spectral subtraction noise reduction method. Of all considered standard measures, a mean squared distortion measure, a SNR based method, the PESQ measure, and the STOI measure yield the highest correlations with the listening experiments scores. These results confirm the ability of instrumental measures to predict the amount of musical noise, but further evaluation shows limitations to their applicability as the results suggest that optimization of the over-subtraction parameter for a minimum amount of musical noise and maximal speech-quality or intelligibly simultaneously, is not possible. Instead the results show that maximal speech-quality or intelligibility is obtained when a stimulus contains the highest amount of musical noise. To gain more insight of the amount of musical noise in a stimulus, a novel measure, based on the characteristics of musical noise in time and frequency, is proposed. This measure incorporates a parametric outlier detection method to classify musical components. High correlations with the outcome of the listening experiment are obtained, i.e. $\rho \approx 0.90$ for enhanced noisy speech signals with various input SNR.

A part of this work has previously been published in [16].

keywords. Instrumental measure, kurtosis, listening experiment, outlier detection, musical noise.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Removing the noise from a noisy speech signal requires the use of a speech enhancement algorithm. Their application can be widely found in fields like telephony or hearing aids where, the goal is to reduce noise without introducing loss of speech-quality or speech-intelligibility, see e.g., [19, 37, 41]. In general, the performance of a speech enhancement algorithm is a trade-off between noise reduction and the extent to which the algorithm distorts the speech signal. Finding the right trade-off is crucial in e.g., 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, only a subjective listening experiment can supply us with the absolute truth about the speech-quality or speech-intelligibility of an enhanced noisy speech signal. Supported by the demand to replace these time consuming and expensive listening tests and, more importantly, to guide algorithm development, there has been a lot of interest in the prediction of speech-quality and speech-intelligibility. This has led to the development of many instrumental speech-quality and speech-intelligibility measures. As these predictors are not perfect, both instrumental measures as well as listening experiments, are essential to evaluate the performance of speech enhancement algorithms. Notice that speech-quality and speech-intelligibility are not the same, there it is possible to have a signal which has perfect speech-intelligibility while having very low speech-quality, therefore, one usually distinguishes between them in algorithm development.

Removing all the noise in a noisy speech signal without introducing loss of speech-quality is unrealistic for any practical application, as speech and noise realizations are unknown. Therefore, many noise reduction methods generally 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 [7], which can remain after processing, and decrease the quality and 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 [7, 9], this is not straightforward without a clear instrumental measure.

While some instrumental measures have shown to predict the quality or intelligibility of enhanced noisy speech with high correlation (e.g. see [21, 48, 50]), 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 instrumental measures to predict the amount of musical noise in the enhanced signal. Given that there

exist 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 2009, Uemura et. al [54] proposed a novel measure based on the kurtosis of the power spectral density (PSD) of a signal, which can be used specifically for predicting the amount of perceived musical noise in the enhanced speech signal. In general, kurtosis provides a measure of the shape of the underlying probability distribution of a random variable, by linking a numerical value to the tailedness and the peakedness. This instrumental measure determines the ratio between the kurtosis of the PSD of an enhanced speech signal and the kurtosis of the PSD of the unprocessed speech signal. It is speculated that by using this higher order statistical method the amount of isolated spectral components can be quantified. Several contributions have been made based upon this metric; e.g. see [22, 23, 58, 45, 46, 51, 52].

Besides this musical noise measure, we consider various other established instrumental measures in terms of their ability to predict the outcome of a listening experiment. Among these are several spectral distance measures, various signal to noise ratio (SNR) based measures, the Perceptual Evaluation of Speech Quality [4] and, the Short-Time Objective Intelligibility Measure [49]. We will imply the descriptions of these predictors as provided in [37, 44, 48, 49]. All these instrumental measures require knowledge of the clean speech signal in order to predict speech-quality or intelligibility. When the clean speech signal needed for this prediction, the measure is referred to as an intrusive measure. Hence, stand-alone measures which are able to predict the speech quality or intelligibility by implying solely the enhanced speech signal are called non-intrusive measures. Most often, only the enhanced noisy speech signal is available at a receivers side, hence only non-intrusive measures are suitable to measure the speech-quality or speech-intelligibility.

In general, the objective of a listening experiment is to determine the quality of a stimulus. To do so, standardized tests like the Mean Opinion Score (MOS) test [29, 26, 25], or the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [27] can be used. However, even though [26] provides a test specifically designed to investigate the intrusiveness of the background noise, it will not capture purely the musicallity of the background distortion, and as a consequence, none of these standard subjective tests provide a suitable method to acquire a description of the amount of musical noise present in a stimulus. Therefore, a listening experiment is designed especially to obtain this description. In total, 80 enhanced noisy speech signals have been investigated and 25 listeners have participated in the process.

The results of this listening experiment characterize the quantity of musical noise in several enhanced noisy speech signals generated by applying the spectral subtraction (SS) noise reduction method [8], and varying the amount of noise reduction. Since SS was first proposed over 30 years ago, research has come up with many adaptations to the orginal algorithm, see e.g., [7, 34]. In this thesis, we will solely focus on how musical noise is produced when applying SS, as the production of musical noise due to SS can be more easily analyzed than using more advanced noise reduction algorithms like the minimum mean square error short-time spectral amplitude (MMSE STSA) estimator [13]. An elaborate discussion

on musical noise produced while making use of the MMSE STSA estimator can be found in [9]. Further, one can fiend an elaborate overview of SS based noise reduction methods in, ([37], Ch. 5) as well as an extensive evaluation of MMSE based estimators ([37] Ch. 6 - 7).

Performance measures are commonly used with the purpose of evaluating the prediction performance of an instrumental measure, e.g., see [37, 21, 16, 48, 49, 50]. As two metrics are commonly used to perform evaluation of the instrumental measures, i.e. the Pearson's correlation coefficient and the Kendall rank correlation coefficient [47], these will be incorporated to examine the performance of the various instrumental measures in terms of their ability to predict the amount of musical noise.

In [16] we showed that the kurtosis based musical noise metric will not provide a well functioning musical noise predictor. However, we did show that certain established instrumental measures are able to predict the outcome of the listening experiment with high correlation. The results indicated that high speech quality or intelligibility is obtained if the amount of musical noise is maximal. Consequently, it becomes impossible to optimize the over-subtraction parameter for both the minimal amount of musical noise and maximal speech quality or intelligibility, at the same time.

To gain better insight in the amount of musical noise in a stimulus, without taking the speech quality or speech intelligibility into account, we propose a new musical noise predictor. This method is based on the characteristics of musical noise in the power spectral domain. Each spectral component is classified based on its observed spectral value and that of its neighbours. This is a procedure commonly performed in postprocessing techniques like time-frequency filtering or median filtering [59, 57, 36], where the spectral components detected as musical components are made inaudible by altering the energy in these components. Moreover, a post-processing method as proposed in [6] utilizes the psychoacoustic criteria in order to the reduce residual noise.

We will view the classification of the spectral components responsible for musical noise in the light of an outlier detection problem. Hawkins [17] defined an outlier as "*an observation that deviates so much from the other observations as to arouse suspicion that if it was generated by a different mechanism*" and, similarly, Johnson [32] defined an outlier as "*an observation in the data set which appears to be inconsistent with the remainder of that set of data*". One can divide outlier detection methods between a class of parametric methods and a class of non-parametric methods. The class of parametric methods require knowledge on the underlying distribution of the data set, where the class of non-parametric methods does not. A large amount of research has been performed on the detection of outliers, however, an elaborate analysis of all possible methods lays beyond the scope of this contribution. For an extensive survey of outlier detection methods we refer to [3, 20, 5]. Based on an estimate of the probability density function (PDF) of the enhanced noisy speech signal, we will explore two, relatively simple, parametric methods, which result in an intrusive and a non-intrusive musical noise measure.

In the finalizing stages of this thesis, we came across another musical noise measure [12]. Analysis of this method is required, however will remain future work.

This thesis is organized as follows. In Chapter 2 we will describe some basic principles of single channel speech enhancement, as well as the production of musical noise and its properties. In Section 2.5, we provide a description of the probability density function (PDF) of an enhanced speech signal, which is the foundation for the method proposed in Chapter 6. After which, the conducted listening experiment is described in Chapter 3. Then, prior to Chapter 5 and 6 in which the considered standard and the proposed instrumental measures are described, the evaluation procedure will be discussed in Chapter 4. In Chapter 7 the performance of the proposed instrumental measures in evaluated using the outcome of the subjective listening experiment.

2. SINGLE CHANNEL SPEECH ENHANCEMENT

2.1 General Principles

Noise reduction in speech recorded by a single microphone is known as single channel speech enhancement (SCSE). In general the noisy speech signal is modeled as a clean speech signal $s[n]$ degraded by an additive uncorrelated noise source $v[n]$, i.e,

$$y_{\text{SNR}}[n] = s[n] + cv[n] \tag{2.1}$$

where $y_{\text{SNR}}[n]$ represents the observed noisy speech signal with time-sample index n. Each of the three signals are commonly considered to be realizations of stochastic processes, moreover, the noise signal, $v[n]$, and the clean speech signal, $s[n]$, are assumed to be statistically independent. Assuming that $v[n]$ has unit-variance, the parameter c denotes a scalar gain regulating the signal-to-noise ratio (SNR). Unless otherwise specified we will denote as $y_{\text{SNR}}[n] = y[n]$ and assume $c = 1$, in this case Eq. (2.1) describes the standard additive noise model [37]. In later chapters the SNR of $y[n]$ will be referred to as the input SNR.

From a statistical perspective, the problem of reducing the noise, $v[n]$, from a noisy observation, $y[n]$ boils down to finding an estimate $\widehat{s}[n]$ of the clean speech sample $s[n]$. In practice, the functionality of speech enhancement methods differs in the domains in which the estimation and processing is performed and, in the (prior) assumptions made on the target or noise processes. Many methods for enhancement of noisy speech have been proposed in literature, e.g. perception based methods, methods in time domain or methods based on hidden markov models. Moreover, a large section of noise reduction algorithms are frequency domain based methods, i.e. subspace based methods that employ the Karhunen-Loeve transform, e.g., [14, 30, 31] or discrete Fourier Transform (DFT) based methods, e.g., [8, 7, 13, 39]. In comparison, the class of DFT based methods offer a good complexity vs. performance trade-off [19], since they are computationally simpler and provide similar overall performance compared to the subspace based methods. An overview of DFT- based noise reduction methods is provided by Hendriks et al. [19].

2.2 Analysis-Modification-Synthesis System

Generally it is assumed that both $s[n]$ and $v[n]$, are wide-sense stationary (WSS). A well known and computationally low cost noise reduction method to exploit

this property is to apply a DFT-based analysis-modification-synthesis procedure [19]. First, a short-time Fourier Transform (STFT) is applied to the noisy speech signal. That is, the noisy speech, $y[n]$, is divided into M overlapping frames of length L with overlap P, after which each frame is multiplied with an analysis window, $w[n]$, and transformed into the DFT domain. Given L to be sufficiently long and the overlap to be sufficiently small, the DFT coefficients can be assumed independent across time and frequency. The STFT coefficients are given as,

$$Y(i,k) = \sum_{n=(L-P)i}^{(L-P)i+L-1} y[n]w[n-(L-P)i] \exp\left(\frac{-j2\pi kn}{L}\right),$$

with $i \in \{0,1,2,...\}$. This results in the following description of the noisy speech model in the DFT domain,

$$Y(i,k) = S(i,k) + V(i,k), \tag{2.2}$$

where, $V(i,k)$, $Y(i,k)$ and $S(i,k)$ denote respectively the zero-mean complex valued random variables of the noise, noisy and target speech processes. The frame-index and transform coefficients are denoted by i and k, respectively. Frames are then processed to reduce noise, which is performed by multiplying the noisy speech coefficients with an appropriate scalar gain, $G(i,k)$,

$$\widehat{S}(i,k) = G(i,k)Y(i,k). \tag{2.3}$$

Afterwards, the enhanced frames, $\widehat{S}(i,k)$, are transformed back into the time domain using an inverse DFT (IDFT). The speech signal is then reconstructed, by applying a synthesis window to each time-frame and using an overlap-add procedure.

2.3 Speech DFT Estimators

One of the first methods proposed to reduce noise by modifying the frequency spectrum of a noisy frame is called the spectral subtraction noise reduction method [8]. SS tries to estimate the spectrum of the clean speech signal by subtracting an estimate of the noise spectrum from the spectrum of the noisy speech process [8, 7, 35, 40].

However, SS is a somewhat heuristically motivated approach. Another downside of SS is the fact that it solely uses information about the PDF of $y[n]$. It is straightforward that by taking into account prior knowledge on the statistical properties of $s[n]$ and $v[n]$ a better clean speech estimate, $\widehat{s}[n]$, can be found. Often, more sophisticated estimators based on minimizing the mean squared error are used to find a statistically optimal representation of the clean speech DFT coefficients.

Most DFT based enhancement methods find a representation for $G(i,k)$ in magnitude or power spectral domain. In the latter case an estimate of the power spectrum can be found by using the periodogram of a time-frame, that is,

$$\begin{aligned} |Y(i,k)|^2 &= |S(i,k) + V(i,k)|^2 \\ &= |S(i,k)|^2 + |V(i,k)|^2 + S(i,k)^* V(i,k) + S(i,k) V(i,k)^*. \end{aligned} \tag{2.4}$$

Generally the periodograms of the noise and target speech processes are unknown and as a consequence, the expected value of a periodogram is often used to provide an estimate of $G(i,k)$. The expected value of a periodogram is better known as the power spectral density. Here, the value of the cross-spectral densities, will be equal to zero, since $s[n]$ and $v[n]$ were assumed to be uncorrelated, that is,

$$\begin{aligned} E[|Y(i,k)|^2] &= E[|S(i,k)|^2] + E[|V(i,k)|^2] \\ &\quad + E[S(i,k)^* V(i,k)] + E[S(i,k) V(i,k)^*] \\ &= E[|S(i,k)|^2] + E[|V(i,k)|^2], \end{aligned} \tag{2.5a}$$

where $E[\cdot]$ denotes the expected value.

Let $R(i,k) = |Y(i,k)|$, $A(i,k) = |S(i,k)|$ and $W(i,k) = |V(i,k)|$ denote the random variables of the magnitude spectrum, i.e the magnitude DFT coefficients (MDFT), of the noisy, clean and noise process, respectively. Moreover, $R^2(i,k)$, $A^2(i,k)$ and $W^2(i,k)$ denote denote the random variables in the power spectral domain (PDFT).

2.4 Power Spectral Subtraction

SS was first proposed by Boll [8] in 1979. Although SS is not theoretically well-founded, it is often used due to its simplicity. The motivation behind SS is based on the fact that the clean speech PSD, $E[A^2(i,k)]$, can be obtained by subtracting the PSD of the additive noise, $E[W^2(i,k)]$, from the PSD of the noisy observation, $E[R^2(i,k)]$, that is,

$$\begin{aligned} E[A^2(i,k)] &= E[R^2(i,k)] - E[W^2(i,k)] \\ &= E[A^2(i,k)] + E[W^2(i,k)] - E[W^2(i,k)]. \end{aligned} \tag{2.6}$$

In practice, the method described by Eq. (2.6), is not realisable as $E[W^2(i,k)]$ is unknown. However, The noise PSD can be estimated by time-averaging noise only periodograms of $R^2(i,k)$, e.g. during speech pauses. The procedure of time-averaging periodograms is better known as Welch's method [56], here the time-average operator is denoted by $\overline{[\cdot]}$. Figure 2.1 depicts such a noisy periodogram and the estimated noise PSD.

An estimate of the clean speech periodogram, $\widehat{A}^2(k)$, can be found by subtracting the noise PSD estimate, $\overline{W^2(k)}$, from a noisy periodogram, $R^2(i,k)$,

$$\widehat{A}^2(i,k) = R^2(i,k) - \overline{W^2(k)}. \tag{2.7}$$

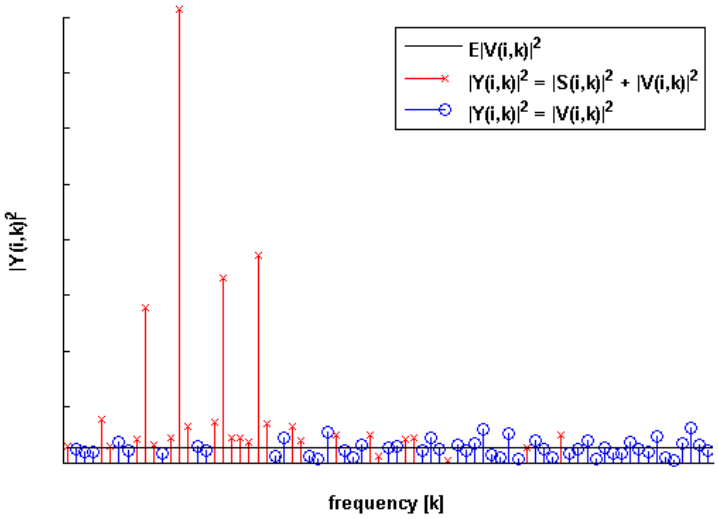


Fig. 2.1: Noisy periodogram and estimated noise PSD

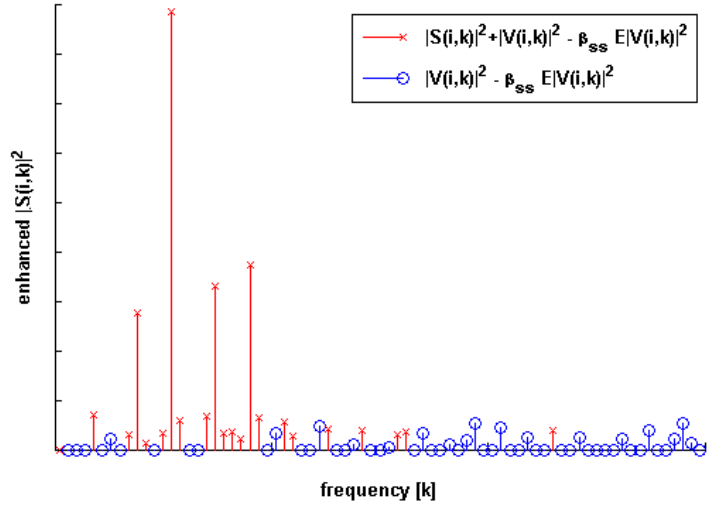
The final estimate $\widehat{S}(i,k)$ is then given by computing $\sqrt{\widehat{A}^2(i,k)}$ and appending the noisy phase.

On account of the variance of the instantaneous noise periodograms, negative spectral values in $\widehat{A}^2(i,k)$ can occur after SS. However, as a PSD is always positive, negative values are overcome by making use of a half-wave rectifier. Another aspect is the variance of the instantaneous noise periodogram, $\overline{W^2(i,k)}$, which results in noise energy that remains present in the enhanced signal. Berouti et al. [7], proposed a modification to reduce the residual noise using a method based on over-subtraction and noise masking,

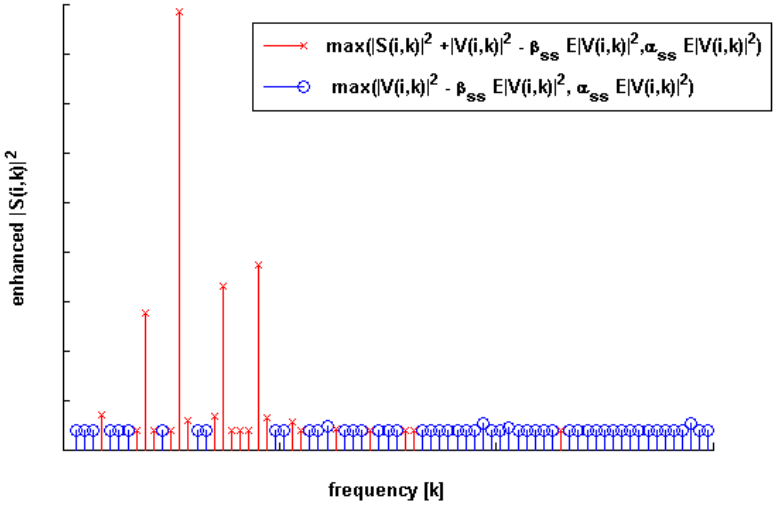
$$\widehat{A}^2(i,k) = \max(R^2(i,k) - \beta_{ss}\overline{W^2(k)}, \alpha_{ss}\overline{W^2(k)}) \tag{2.8}$$

where, β_{ss} is called the over-subtraction factor and represents a scalar weight on $\overline{W^2(i,k)}$ and, α_{ss} denotes a flooring parameter which introduces a minimum value on $\widehat{A}^2(i,k)$.

A more illustrative example of the effect of SS (Eq. (2.8)) on a noisy speech periodogram is provided in Fig. 2.2(a) and Fig. 2.2(b). Comparing Fig. 2.1 and Fig. 2.2(a) shows that several noise realizations have been eliminated, although, it may also be observed that $\widehat{A}^2(i,k)$, in Fig. 2.2(a), still contains noise energy. The amount of narrowband residual noise will decrease if β_{ss} is increased, however, increasing β_{ss} reduces the signal power in every frequency bin and, consequently speech energy will also be eliminated, resulting in a distorted clean speech estimate. Noise masking, i.e. $\alpha_{ss} > 0$, allows broadband noise to remain present in an attempt to mask the narrow band residual noise, as broadband noise is often perceived less annoying as the residual noise. This latter procedure is depicted in Fig. 2.2(b). Generally, altering β_{ss} and α_{ss} provides a trade-off



(a) $\beta_{ss} = 2, \alpha_{ss} = 0$



(b) $\beta_{ss} = 2, \alpha_{ss} \in \mathbb{R}^+$

Fig. 2.2: The effect of using oversubscription and residual noise masking

between the amount of narrow-band residual noise, broadband noise reduction and speech distortion.

2.5 Power Spectral Subtraction In Terms Of PDF's

There has been a lot of discussion on how to model the underlying probability distribution of $S(i,k)$ and $V(i,k)$, [19]. If it is assumed that the noise source consists of the sum of multiple independent noise sources, then the central limit

theorem (CLT) ensures that the distribution of the recorded noise process in the DFT domain, will approaches a complex gaussian distribution. Moreover, if the time-span of dependency between the samples of the observed process is also small, then even for a single source, the distribution in the will approach a complex gaussian distribution in the DFT domain [19]. Therefore, let us assume, that both $S(i,k)$ and $V(i,k)$ follow a zero-mean complex Gaussian distribution, $S(i,k) \sim \mathcal{CN}(0,\sigma_S^2)$ and $V(i,k) \sim \mathcal{CN}(0,\sigma_V^2)$. Here σ_S^2 and σ_V^2 describe the variance of the complex valued speech and noise DFT coefficients. The PDF of the noisy DFT coefficient $Y(i,k)$, can now be written as the PDF of the sum of the two independent random variables $S(i,k)$ and $V(i,k)$, i.e., as the convolution of their separate probability density functions. Consequently, the sum of two gaussian PDFs will again provide a Gaussian distribution. The underlying PDF of the noisy speech process can then be derived as follows,

$$p_Y(y) = p_{S+V}(y) = (p_S * p_V)(y) \tag{2.9a}$$

$$Y(i,k) \sim \mathcal{CN}(0,\sigma_S^2+\sigma_V^2) = \mathcal{CN}(0,\sigma_Y^2), \tag{2.9b}$$

where $\sigma_Y^2 = \sigma_S^2 + \sigma_V^2$ denotes the variance of the complex valued noisy speech DFT coefficients. As described in Chapter 2.1, the noisy speech signal is first transformed into the MDFT domain and is denoted by $R(i,k)$. The underlying PDF of $R(i,k)$, can be described by a Rayleigh distribution [19], where the variance consists of the joint variance of the real and imaginary part of the DFT coefficients, $\sigma_Y^2 = \sigma_{\mathrm{Re}(Y)}^2 + \sigma_{\mathrm{Im}(Y)}^2$,

$$p_R(r) = \begin{cases} \frac{2r}{\sigma_Y^2}\exp\left(-\frac{r^2}{\sigma_Y^2}\right) & r \geq 0 \\ 0 & otherwise. \end{cases} \tag{2.10a}$$

The MDFT coefficients are then transformed into the power spectral domain, where their underlying PDF can be described by the following exponential distribution [18],

$$p_{R^2}(r^2) = \begin{cases} \frac{1}{\sigma_Y^2}\exp\left(-\frac{r^2}{\sigma_Y^2}\right) & r^2 \geq 0 \\ 0 & otherwise \end{cases} \tag{2.11a}$$

where $E[R^2] = \sigma_Y^2$ and $Var[R^2] = \left[\sigma_Y^2\right]^2$.

An estimate of the clean speech signal, $\widehat{A}^2(i,k)$, can be found by applying the spectral subtraction method discussed in Chapter 2.4. SS consists of two consecutive steps, first, the PDF is shifted to the left due to subtraction of $\beta_{ss}E[W^2]$. The shifted PDF can be described by,

$$p_{R^2}(r^2) = \begin{cases} \frac{1}{\sigma_Y^2}\exp\left(-\frac{r^2+\beta_{ss}E[W^2]}{\sigma_Y^2}\right) & r^2 > \beta_{ss}E[W^2] \\ 0 & otherwise \end{cases} \tag{2.12}$$

The half-wave rectifier then ensures all coefficients, which have obtained negative spectral values during the first step, to be located at position zero. The probability mass of these negative valued coefficients, $Pr_{zero} = Pr(r^2 - \beta_{ss}E[W^2] < 0)$, is clustered and can be modeled by a δ-function located at zero. Pr_{zero} can be calculated as follows,

$$\begin{aligned} Pr_{zero} &= \int_0^{\beta_{ss}E[W^2]} \frac{1}{\sigma_Y^2} \exp\left(-\frac{t^2}{\sigma_Y^2}\right) dt \\ &= 1 - \exp\left(-\frac{\beta_{ss}E[W^2]}{\sigma_Y^2}\right). \end{aligned} \tag{2.13}$$

Consequently, the PDF of the enhanced signal $\widehat{A}^2(i,k)$, as function of noisy power spectral components, r^2, can be denoted by,

$$p_{\widehat{A}^2}(r^2) = \begin{cases} \frac{1}{\sigma_Y^2} \exp\left(-\frac{r^2+\beta_{ss}E[W^2]}{\sigma_Y^2}\right) + \left(1 - \exp\left(-\frac{\beta_{ss}E[W^2]}{\sigma_Y^2}\right)\right)\delta[r^2] & r^2 \geq 0 \\ 0 & otherwise \end{cases} \tag{2.14}$$

where the last term in Eq. (2.14), i.e. $1 - \exp(\cdot)$, represents the probability mass of the spectral components which have been set to zero.

Subsequently, both the expected value and the variance of $\widehat{A}^2(i,k)$ can be obtained using Eq. 2.14 and are , respectively, denoted by,

$$E[\widehat{A}^2] = \sigma_Y^2 \exp\left(-\frac{\beta_{ss}E[W^2]}{\sigma_Y^2}\right), \tag{2.15a}$$

$$Var[\widehat{A}^2] = \left[\sigma_Y^2\right]^2 \left(2\exp\left(-\frac{\beta_{ss}E[W^2]}{\sigma_Y^2}\right) - \exp\left(-\frac{2\beta_{ss}E[W^2]}{\sigma_Y^2}\right)\right). \tag{2.15b}$$

2.6 Musical Noise Produced By Spectral Subtraction

The origin of musical noise can be found in estimating the noise periodogram, $W^2(i,k)$. Namely, 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. This problem becomes even worse, as the PSD is unknown and replaced by an estimate that can have a high variance. Due to the variance of the instantaneous noise periodogram it is possible that noise energy will remain present in the enhanced speech periodogram, i.e. when $W^2(i,k) > \beta_{ss}E[W^2]$.

A more vivid visualization of this phenomenon can be obtained when considering several consecutive time-frames of an enhanced noise only signal, as shown in Fig. 2.3, or by observing the spectrogram of an enhanced noisy speech signal, as depicted in Fig. 2.4.

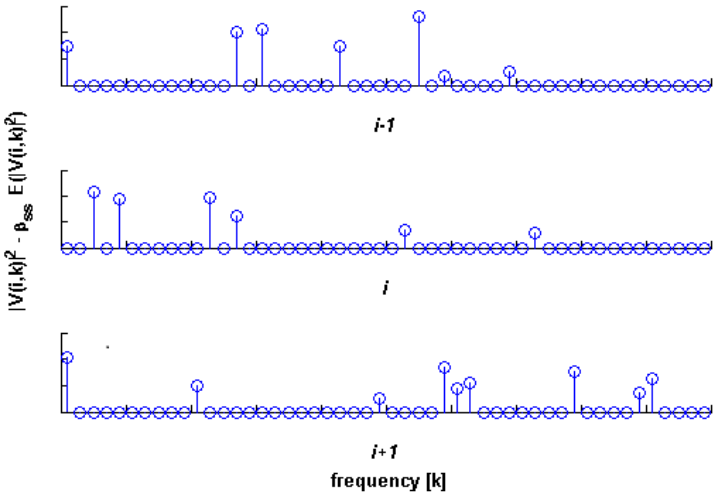


Fig. 2.3: Residual noise responsible for musical noise in three consecutive noise only time-frames of $|\widehat{S}|^2$.

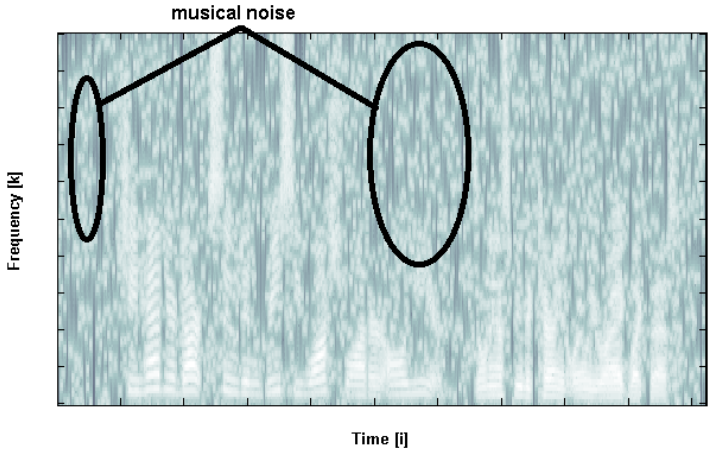


Fig. 2.4: Spectrogram of an enhanced speech signal containing musical noise, the energy within the circles can be classified as isolated spectral peaks which may produce a musical artifacts

As a consequence the noise variance, spectral components in $\widehat{A}^2(i,k)$ can become isolated in time and frequency, i.e., a spectral component can have a non-zero energy in a zero energy time-frequency region. The power residue will introduce a short tonal artifact after reconstruction of the time domain signal. This is called a musical artifact. When multiple isolated spectral peaks occur in various time-frames and across different frequency bands, then we define the collection of the musical artifacts as the *number* of musical artifacts within the enhanced noisy speech signal.

As will be argued the following paragraph, it can be concluded that the number of musical artifacts occurring in an enhanced noisy speech signal, is in fact is independent of the SNR of the noisy speech signal.

Consider a noisy speech signal, $y_{\text{SNR}}[n]$, as defined in Eq. (2.1) which is transformed into power spectral domain. Then, an estimate of the clean speech periodogram can be found by applying SS, Eq. (2.8),

$$\begin{aligned}\widehat{A}^2(i,k) &= \max(A(i,k)^2 + 2cA(i,k)W(i,k) + c^2W^2(i,k) - \beta_{ss}c^2E[W^2], 0) \\ &= \max(A(i,k)^2 + c^2W^2(i,k) - \beta_{ss}c^2E[W^2], 0).\end{aligned} \tag{2.16}$$

Most of the audible musical artifacts occur in noise-only time-frequency bins. Therefore, when assuming noise-only bins, Eq. (2.16) reduces to,

$$\begin{aligned}\widehat{A}^2(i,k) &= \max(c^2W^2(i,k) - \beta_{ss}c^2E[W^2], 0) \\ &= \max(c^2(W^2(i,k) - \beta_{ss}E[W^2]), 0)\end{aligned} \tag{2.17}$$

A residual noise component will remain in the enhanced signal if $c^2W^2(i,k) > c^2\beta_{ss}E[W^2]$ and, the scaling parameter c will scale the energy of this residual noise component. Thus, if $W^2(i,k)$ represents an isolated spectral peak, scaling the SNR of $y_{\text{SNR}}[n]$ by means of c will only scale the energy of this musical artifact and it will have no effect on the number of musical artifacts.

As it is known that the loudness of musical noise is sensitive to scaling of the SNR of $y_{\text{SNR}}[n]$, we can define the amount of musical noise in an enhanced noisy speech signal as follows,

The amount of musical noise in a stimulus
The amount of musical noise can be defined as the cumulative sum of the energies of the isolated spectral peaks in a stimulus, where an isolated spectral peak induces a musical artifact with a certain loudness.

It is often assumed that musical artifacts only appear due to enhanced noise-only time-frequency components, however, this assumption is only valid provided the amount of noise reduction is much larger than the amount of speech distortion. That is, certain values of β_{ss} can introduce large amounts of speech distortion, and consequently also a speech component can become isolated in time and frequency which may result in a short tonal sound.

Note that from a perceptual point of view not all isolated spectral components will produce an audible musical artifact, due to the masking properties of the human hearing. Psychoacoustic models which exploit efficiently these masking characteristics can be used to determine whenever an isolated spectral component is audible. For an elaborate description of a psychoacoustic model see e.g. [43]. We will assume that the amount of isolated spectral peaks is equal to the number of audible musical artifacts.

3. SUBJECTIVE MEASUREMENTS

When standardized subjective methods do not suffice, a listening experiment which is explicitly developed with the purpose of determining the amount of musical noise in a stimulus must be designed. This chapter describes the listening experiment that was conducted to be able to evaluate the performance of the existing and proposed predictors of musical noise.

3.1 Description Of The Stimuli

A noisy speech signal is obtained by degrading a clean speech signal, $s[n]$, with additive white Gaussian noise. For the conducted experiment, a set of four noisy signals have been obtained using Eq. (2.1), having an input SNRs of, 0 dB, 5 dB, 10 dB and 15 dB, respectively. That is,

$$\{y_{\text{SNR}}[n]\} = \{y_{0\ \text{dB}}[n], y_{5\ \text{dB}}[n], y_{10\ \text{dB}}[n], y_{15\ \text{dB}}[n]\}.$$

The clean speech signal which was used, $s[n]$, consisted of a concatenation of two sentences [15], which were read by a female and a male talker, respectively.

source	[15]
$s[n]$	SA1, SA2
$v[n]$	$\mathcal{N}(\mu_v, \sigma_v^2)$
SNR	0, 5, 10, 15 dB
length	7 sec
Fs	16 kHz
$w[n]$	square-root Hann window
overlap	50 %
L	32 msec
NRM	SS (Eq. (2.8))
β_{ss}	{0, 0.5, ... , 9.5}
α_{ss}	0

Tab. 3.1: Details on the investigated enhanced noisy speech signals

All noise reduction was performed in the power spectral domain, using time frames of 32 ms with 50% overlap, a 512-point FFT was used to transform each time-frame into the DFT domain and, a square-root Hann window is used as analysis and synthesis window. Further, the noise PSD was estimated over 10 seconds of noise-only signal. Note that this latter noise process is not equivalent

to the noise processed used to generate the noisy speech signal, however, it follows an identical PDF.

The clean speech estimates, $\{\widehat{s}_{\text{SNR, }\beta_{ss}}[n]\}$, are generated by applying SS, Eq. (2.8), with various over-subtraction factors β_{ss} to $\{y_{\text{SNR}}[n]\}$. Since altering the over-subtraction factor produces enhanced speech signals which contain different amounts of musical noise. A suitable range is found by varying β_{ss} in steps of 0.5, starting from $\beta_{ss} = 0$, where there is no noise reduction, to $\beta_{ss} = 9.5$, where the enhanced noisy speech signal consists of highly distorted speech but contains little to no musical noise. Note that such a large value of β_{ss} is rarely used in practice, due to the amount of speech distortion it introduces. As masking of residual background noise is undesired for this specific experiment, the spectral floor, α_{ss}, is set to 0.

By applying this procedure to the four noisy speech signals, $\{y_{\text{SNR}}[n]\}$, a set of 80 enhanced noisy speech signals is obtained,

$$\begin{aligned}\{\widehat{s}_{\text{SNR, }\beta_{ss}}[n]\} = \{&\widehat{s}_{0\text{ dB, }0}[n], \widehat{s}_{0\text{ dB, }0.5}[n], \\ &..., \widehat{s}_{0\text{ dB, }9.5}[n], \widehat{s}_{5\text{ dB, }0}[n], ..., \widehat{s}_{15\text{ dB, }9.5}[n]\}.\end{aligned} \tag{3.1}$$

Prior to presenting the enhanced signals to the listeners, the set is randomized.

3.2 Listening Experiments

The listening experiment is designed specifically to characterize the amount of musical noise in an enhanced noisy speech signal. Listeners were asked to 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?'.

The signals are to be 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. 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 shown in Table 3.2.

score	**The Mean-Opinion Musical Noise Score (MOMN-score)**
5	Extreme
4	A lot
3	Medium
2	A little
1	Broadband noise or no musical noise

Tab. 3.2: Grading scale for the amount of musical noise

Comparing a set of 80 signals at once is too tiring for a listener, therefore the set is split into eight different subsets containing 10 stimuli each. The graphical user interface which was presented to the listeners is depicted in Fig. 3.2.

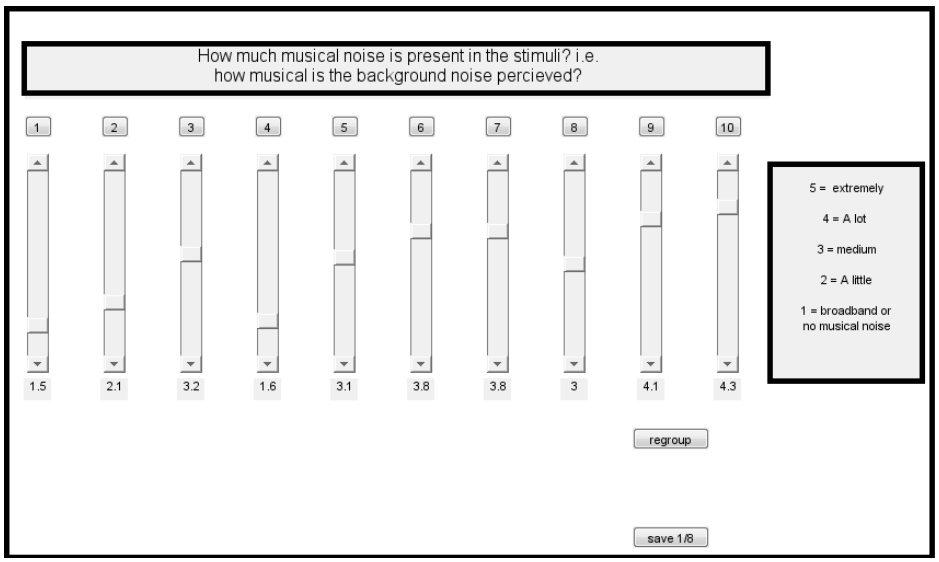


Fig. 3.1: Graphical user interface (GUI) used in the listening experiment

In total, the listening experiment was performed by 25 subjects, 90% male, and 10% female within the age-range of 21 up to 51 years old. Furthermore, the experiments were done inside a sound proof listening room [24]. All listeners were normal hearing to the best of their knowledge.

The mean opinion musical noise score is calculated by averaging over the obtained scores for each over-subtraction parameter β_{ss}. These average MOMN-scores for the four different input SNRs are depicted Fig. 3.2(a).

3.3 Discussion Of The Results

In Chapter 2.6 it was concluded that the number of musical artifacts does not change if the noise variance is scaled, however, the energy of these musical artifacts does change. From Fig. 3.2(a) it can be seen that the perceived amount of musical noise varies with SNR, hence, the energy of the musical artifacts contributes to the amount of musical noise perceived, which could be expected as energy and loudness is positively correlated. This confirms the definition of the amount of musical provided in Chapter 2.6.

Upon reviewing the results of the listening experiment in Fig. 3.2(a), the transition from a noisy speech signal to a nearly clean, but heavily distorted, speech signal in terms of the amount of perceived musical noise can be observed. That is, in the case of $\beta_{ss} = 0$, the noise distortion consists purely of the added noise source. Then, for larger values of β_{ss}, one will observe a transition from broadband noise into large amounts of musical noise. It can be seen that for $\beta_{ss} \approx 2.5$ the amount of musical noise is maximal, where for larger values of β_{ss} the amount of musical noise reduces until nearly none is left.

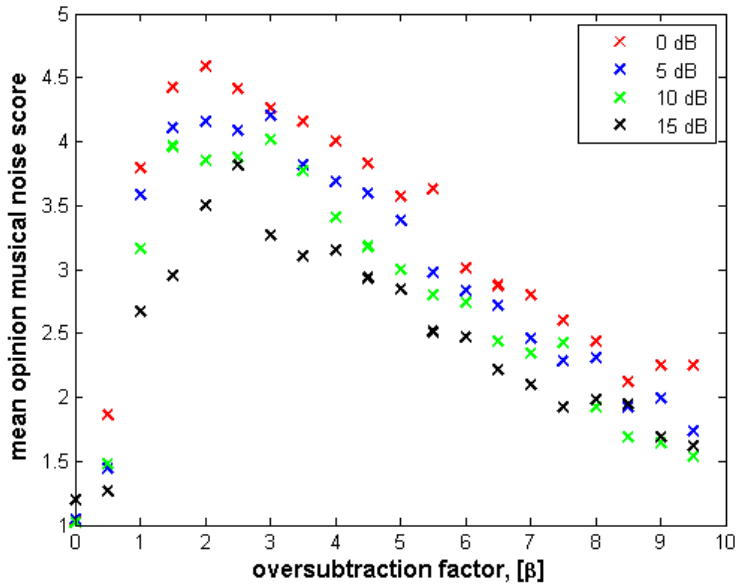


Fig. 3.2: Results obtained from 25 listening experiments. Each x is an average score for a particular { SNR β }-pair across the 25 listeners

4. EVALUATION PROCEDURE

The listening experiment performed in Chapter 3 provides us with a description of the amount of musical noise as a function of the MOMN-score. To find an instrumental measure which is able to predict the amount of musical noise, all considered instrumental measures will be evaluated using data-set $\{\widehat{s}_{\text{SNR, }\beta_{ss}}[n]\}$, as described in Chapter 3. We want to measure to which extent a predictor correlates with the outcome of each listening experiment, i.e., we wish to measure the performance of an instrumental measure in terms of its ability to predict the amount of musical noise in a stimulus. Consequently, an instrumental measure which predicts the perceived amount of musical noise in a stimulus is able to predict the outcome of the listening experiment with high correlation.

Two different performance metrics are applied. The first measure, the Pearson correlation coefficient, is denoted by ρ, see e.g. [47],

$$\rho = \frac{\sum_l (S_l - \overline{S})(D_l - \overline{D})}{\sqrt{\sum_l (S_l - \overline{S})^2 \sum_l (D_l - \overline{D})^2}}. \tag{4.1}$$

where S and D describe the outcome of the listening experiment and the scores from the instrumental measure, respectively. $\overline{S}$ and $\overline{D}$ define the expected value of the sets S and D, and l denotes the over-subtraction index. Note that ρ is limited between -1 and 1. A high correlation coefficient $|\rho|$ yields a good performing musical noise metric. Additionally, the Kendall's tau rank correlation coefficient τ, is used to support any conclusions drawn based on the Pearson correlation coefficient, it is defined as,

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}, \tag{4.2}$$

where, N_c describes the concordant pairs, while N_d describes the discordant pairs [47]. N denotes the total number of enhanced speech signals applied for the subjective test, which will be equal to $N = 25$. As for the Pearson correlation coefficient, τ is limited between -1 and 1.

Both performance measures ρ and τ, will provide a metric which evaluates a linear relationship between the outcome obtained in the listening experiment and the results from the instrumental measure.

5. INSTRUMENTAL MEASURES

In this chapter, we examine various established instrumental measures which are generally used as quality or intelligibility measures in the field of speech coding and speech enhancement. A short description of these measures and their applicability is provided in section 5.1 and 5.2. Table 5.1 provides an overview of the considered metrics, where the implementations which are used in obtaining the results in chapter 7 are provided in the most right column.

abriviation	**Instrumental Measure**	**Implementation used**
EUCL	Euclidean Distance [44]	[1]
MSD	Magnitude Spectral Distance [48]	
MSD2	Power Spectral Distance	
LSD	Log-Spectral Distance[48, 37]	[37]
LLR	Log-Likelyhood ratio [44, 33]	[37]
IS	Itakura-Saito Distance [44, 37, 28, 1]	[1]
SNRseg	Segmental SNR [44, 37]	[37]
fwSNRseg	Frequency Weighted Segmental SNR [37, 44, 38]	[37]
fwSNRsegn	Normalized Frequency Weighted Segmental SNR [21, 37]	[37]
CEP	Cepstral Distance [44, 33, 48]	[37]
PESQ	Perceptual Evaluation of Speech Quality [42, 4, 37]	[37]
COMPovl	Composite Noise Distortion Measure, overall quality[21, 37]	[37]
COMPbn	Composite Noise Distortion Measure, background distortion[21, 37]	[37]
STOI	Short-Time Objective Intelligibility Measure[49]	[49]
KBMN	Kurtosis Based Musical Noise Measure [54]	

Tab. 5.1: The various different instrumental measures used for evaluation

All these instrumental measures require the clean speech signal, $s[n]$, to be known exactly. The clean speech process can either be used as a reference signal, or to determine a voice activity detector (VAD). Here a VAD is defined as follows,

$$\begin{aligned} \text{Voice present} &: 10\log_{10}(\max(A^2(i,k))) - 10\log_{10}(A^2(i,k)) > \varrho \\ \text{Voice absent} &: 10\log_{10}(A^2(i,k)) - 10\log_{10}(A^2(i,k)) < \varrho, \end{aligned} \tag{5.1}$$

where, threshold ϱ is chosen equal to 50 dB.

5.1 Standard Instrumental Measures

From the predictors listed in Table 5.1, the first seven metrics calculate the distance between the spectra of a clean speech signal, $A(i,k)$, and the enhanced noisy speech signal, $\widehat{A}(i,k)$. These measures are often called "spectral distance measures". The first metric, the Euclidian distance measure (EUCL), can be

denoted by,

$$d_{\text{EUCL}} = \frac{1}{K} \sum_{k} \sqrt{\sum_{i} |A^2(i,k) - \widehat{A}^2(i,k)|^2}, \tag{5.2}$$

where K describes the total number of frequency coefficients, and M denotes the total number of time-frames. The measures, MSD and MSD2, are strongly related to EUCL, but are computed for the MDFT coefficients, and the periodogram of the estimated clean speech signal and the reference signal, respectively,

$$d_{\text{MSD}} = \frac{1}{K} \sum_{k} \sqrt{\frac{1}{L} \sum_{i} |A(i,k) - \widehat{A}(i,k)|^2}, \tag{5.3}$$

$$d_{\text{MSD2}} = \frac{1}{K} \sum_{k} \sqrt{\frac{1}{L} \sum_{i} |A^2(i,k) - \widehat{A}^2(i,k)|^2}. \tag{5.4}$$

Likewise, the log spectral distance measure (LSD), the log-likelihood ratio measure (LLR) and the Itakura-Saito distance measure (IS) are defined as,

$$d_{\text{LSD}} = \frac{1}{K} \sum_{k} \sqrt{\frac{1}{L} \sum_{i} |20\log_{10}(A(i,k)) - 20\log_{10}(\widehat{A}(i,k))|^2}. \tag{5.5}$$

$$d_{\text{LLR}} = \frac{1}{M} \sum_{i} \log\left(1 + \sum_{k} \frac{|A(i,k) - \widehat{A}(i,k)|^2}{|A(i,k)|^2}\right), \tag{5.6}$$

$$d_{\text{IS}} = \frac{1}{M} \sum_{i} \sum_{k} \left[\frac{A^2(i,k)}{\widehat{A}^2(i,k)} - \log_{10}\left(\frac{A^2(i,k)}{\widehat{A}^2(i,k)}\right) - 1\right]. \tag{5.7}$$

Furthermore, three SNR based predictors are considered; the segmental SNR measure (SNRseg), the frequency weighted segmental SNR measure (fwSNRseg) and the normalized frequency weighted segmental SNR measure (fwSNRsegn). These three methods calculate the average of the SNR over short time segments. That is, the time-domain signal is divided into M overlapping time-frames of length L, after which each frame is multiplied by a Hann-window, $w[n]$. The average SNR per time-frame will then result in the segmental SNR,

$$d_{\text{SNRseg}} = \frac{20}{M} \sum_{i} \log_{10}\left(\frac{\sum_{n=Nm}^{Nm+L-1} |s[n]w[n]|}{\sum_{n=Nm}^{Nm+L-1} |s[n]w[n] - \widehat{s}[n]w[n]|}\right). \tag{5.8}$$

The fwSNRseg measure, proposed by Tribolet et al. [53], is an extension of SNRseg, with an additional averaging over frequency bands. Here, the frequency bands are obtained by applying a DFT based critical band decomposition. An adjusted version of fwSNRseg, the fwSNRsegn measure, is proposed by Ma et al. [38], where the values of clean and enhanced noisy speech signal are first normalized between -1 and 1, after which fwSNRseg is calculated as before. Now

note that both the result of SNRseg and fwSNRseg measure are sensitive to scaling of the processed speech signal, however, because of this normalization step prior to calculating fwSNRseg, the fwSNRsegn measure becomes insensitive to scaling of the enhanced noisy speech signal.

As for the cepstral distance measure (CEP) measure, both LLR and IS can be constructed in terms of linear prediction coefficients (LPC). In this case, the measures assume that speech is an autoregressive process for short-time segments which is modeled with a linear prediction method. Moreover, CEP is a function of the cepstral representation of the LPC's. Mathematical details can on CEP, LLR and IS as function of LSP's can be found in [44, 37].

Beerends et al.[4] proposed a speech-quality predictor, the Perceptual Evaluation of Speech Quality measure (PESQ). The PESQ measure provides a prediction of the MOS-score. Because the PESQ measure is too complex to describe elaborately, an extensive discussion of PESQ lays beyond the scope of this contribution. However, Taal et al. [48], provided a short but clear description of this measure; "*First, the clean and processed speech are time aligned in order to compensate for any delay differences, after which both signals are processed by a psycho-acoustical model to obtain their internal representations. After global and local normalization these representations are compared resulting in so-called time-frequency dependent disturbance densities. By combining these values a PESQ-score is obtained.*"

Based on linear combination of several speech-quality predictors, Hu et al. [21] proposed a composite measure for the background noise distortion defined by ([37] p. 572),

$$\begin{aligned} d_{\mathrm{COMPovl}} = 1.227 + 0.334 \cdot \mathrm{fwSNRseg} + 0.347 \cdot \mathrm{PESQ} - 0.682 \cdot \mathrm{LLR} \\ - 0.006 \cdot \mathrm{IS} + 0.141 \cdot \mathrm{CEP} + 0.033 \cdot \mathrm{SNRseg} - 0.107 \cdot \mathrm{WSS}, \end{aligned} \tag{5.9}$$

and an overall quality measure,

$$\begin{aligned} d_{\mathrm{COMPbn}} = 1.001 + 0.318 \cdot \mathrm{fwSNRseg} + 0.533 \cdot \mathrm{PESQ} - 0.852 \cdot \mathrm{LLR} \\ + 0.006 \cdot \mathrm{IS} + 0.143 \cdot \mathrm{CEP} - 0.132 \cdot \mathrm{WSS}. \end{aligned} \tag{5.10}$$

Where WSS denotes the weighted spectral slope measure which computes the weighted difference between the spectral slopes in each frequency band, see ([37] p. 508). There the WSS measure penalizes heavy differences in spectral peak locations [44], also the narrow-band spectral peaks responsible for musical artifacts will be penalized heavily, causing the WSS measure as a stand-alone measure for musical noise prediction to be insufficient.

Although it was concluded in [21] that the measure COMPbn does not provide a good predicting of the noise distortion in a signal, it represents the only measure especially designed to predict the background distortion.

Finaly, the short-time objective intelligibility (STOI) measure [49] is implied. The STOI measure quantifies, in contraire to the other investigated instrumental measure which predict speech-quality, the speech-intelligibility of speech present in a stimulus. Mathematical details on the STOI measure can be found in [49].

5.2 Kurtosis Based Musical Noise Measure

Remark: *Uemura et. al [54] state that the number of isolated spectral peaks is equivalent to the amount of musical noise. This definition is not equivalent to the definition of the amount of musical noise provided in Chapter 2.6. Here we defined the number of isolated spectral peaks to be equivalent to the number of musical artifacts, but the number of musical artifacts is not equivalent to the amount of musical noise.*

The kurtosis based musical noise metric (KBMN) quantifies the number of isolated spectral peaks in an enhanced signal by determining the ratio between the kurtosis of an enhanced and noisy speech signal,

$$\text{kurtR} = \frac{\mathcal{K}_{\widehat{A}^2}}{\mathcal{K}_{R^2}} \tag{5.11}$$

It is stated that a high kurtosis ratio implies an enhanced signal containing a large amount of musical noise.

However, due to the somewhat vague definition of kurtosis in [54], contradictions within these contributions have occurred. In [54] the kurtosis is estimated as,

$$\mathcal{K}_{X^2} = \frac{\mu_4}{\mu_2^2}, \tag{5.12}$$

where μ_n denotes the n-th order moment of a random variable X^2. Here X^2 denotes either the enhanced or noisy speech signal in the power spectral domain, i.e. $\widehat{A}^2(i,k)$ or $R^2(i,k)$. Originally, the kurtosis of a PDF is defined by the standardized fourth order population moment,

$$\mathcal{K}_{X^2} = \frac{E\left[(X^2 - E[X^2])^4\right]}{\left(E\left[(X^2 - E[X^2])^2\right]\right)^2} = \frac{\mu_4}{\sigma^4}, \tag{5.13}$$

where $E[X^2]$ denotes the ensemble average of random variable X^2. An elaborate evaluation of the kurtosis as defined in Eq. (5.13) can be found in [11].

Huanjun et al. [22] use the definition of kurtosis as denoted in Eq. (5.13) to produce an instrumental musical noise measure applicable to signals enhanced with unknown noise reduction methods. Note however, that this implementation of kurtosis is, in general, not equivalent to the definition of kurtosis provided in Eq. (5.12). If reviewing Eq. (5.12) in the light of Eq. (5.13) it can be concluded that the authors of [54] assume $E[X^2] = 0$. However, in power spectral domain the realizations are strictly positive, i.e. the support of the PDF is non-negative. Hence, only in the case of zero signal energy, $E[X^2]$ will be equal to zero.

Yong et al. [58] employ a definition of kurtosis in the DFT domain to compute kurtR. In literature the method of computing the kurtosis of the PDF of a signal in the DFT domain is better known as the spectral kurtosis,

$$\mathcal{K}_X = \frac{E[X^4]}{\left(E[X^2]\right)^2} - 2, \tag{5.14}$$

Here X denotes the MDFT coefficients of the enhanced or noisy speech signal. A proof is provided in [55, 2]. Note that also Eq. (5.14) is not equivalent to Eq. (5.12).

Across all these contributions it is assumed that musical noise solely occurs in noise-only signal segments. To ensure the validity of assumption, Yong et al. [58] determine the noise-only components by using a multi-decision sub-band VAD [10], where other contributions assume to have complete knowledge of the clean speech signal [22, 45, 51, 52, 23].

6. OUTLIER BASED MUSICAL NOISE MEASURE

As was concluded in Chapter 2.6 and Chapter 3, the amount of musical noise perceived is dependent on the number of musical artifacts and the energy of these musical artifacts. In this chapter we will present a measure which estimates the perceived amount of musical noise by first determining an estimate of the number of musical artifacts using a hypothesis-test based procedure. Here we exploit the assumption that spectral components that are responsible for musical noise exhibit an isolated character. The perceived amount of musical noise is then predicted by calculating the sum of the energies of the identified isolated spectral peaks. We will call this measure the outlier based musical noise (OBMN) measure.

6.1 Musical Noise Predictor For An Enhanced Noise-Only Signal

It was argued in Chapter 2.6, that musical noise occurs most often in enhanced noise-only signal segments, i.e, $R^2(i,k) = W^2(i,k)$. Noise energy will remain in the processed signal when $\widehat{A}^2(i,k) = \max(W^2(i,k) - \beta_{ss}E[W^2], 0)$ and $W^2(i,k) > \beta_{ss}E[W^2]$, which may result in an isolated spectral peak responsible for a musical artifact. To exploit this isolated characteristic, consider a $(2Z+1)\times(2P+1)$ sized window, $\widehat{\mathbf{A}}^2$, centered around $\widehat{A}^2(i,k)$, describing the neighbourhood around the enhanced spectral component $\widehat{A}^2(i,k)$.

$$\widehat{\mathbf{A}}^2 = \begin{bmatrix} \widehat{A}^2(i-Z,k-P) & \cdots & \widehat{A}^2(i-Z,k+P) \\ & \ddots & \\ \vdots & \widehat{A}^2(i,k) & \vdots \\ & \ddots & \\ \widehat{A}^2(i+Z,k-P) & \cdots & \widehat{A}^2(i+Z,k+P) \end{bmatrix}^T \tag{6.1}$$

Z and P can be chosen arbitrary, however the frequency and time resolution need to be high enough to allow the assumption of local stationarity within $\widehat{\mathbf{A}}^2$. All spectral components within neighbourhood $\widehat{\mathbf{A}}^2$ can then be considered as values drawn from a same PDF. Hence, it can be concluded that if $\widehat{A}^2(i,k)$ represents an isolated spectral peak, it is in fact an outlier of the distribution described by Eq. (2.14). Classifying whether the spectral value $\widehat{A}^2(i,k)$ is responsible for a isolated spectral peak boils down to an outlier detection problem.

OBMN Measure
If a spectral component $\widehat{A}^2(i,k)$ is larger then a certain threshold η, then $\widehat{A}^2(i,k)$ is classified as an isolated spectral peak responsible for a musical artifact. The threshold η can be calculated using the enhanced noisy neighbourhood $\widehat{\mathbf{A}}^2$ centered around $\widehat{A}^2(i,k)$ as follows,

$$\begin{aligned} &\mathrm{H}_0 : \widehat{A}^2(i,k) \leq \eta \\ &\mathrm{H}_1 : \widehat{A}^2(i,k) > \eta. \end{aligned} \tag{6.2}$$

In Chapter 3 it was concluded that a combination of the number of musical artifacts and the perceptibility of these artifacts is representative for the amount of perceived musical noise. Therefore, the number of detected musical artifacts per frame is defined as the Tempo of the musical noise, where the amount of musical noise is defined as the sum of the power of the detected outliers.

$$Tempo = \frac{|\{\widehat{A}^2(i,k) \in H_1\}|}{M}, \tag{6.3}$$

$$Amount\ of\ musical\ noise = \sum(|\{\widehat{A}^2(i,k) \in H_1\}|), \tag{6.4}$$

where H_1 denotes the set of spectral components classified as outliers and $|\cdot|$ denotes the cardinality of the set.

The problem of predicting the amount of musical noise has become a problem of finding the threshold η. That is, if a spectral component $\widehat{A}^2(i,k)$ is larger then a certain threshold η, then $\widehat{A}^2(i,k)$ will be classified as a non-zero noise component.

6.2 Methods Of Determining η

For a spectral component $\widehat{A}^2(i,k)$, to be an outlier responsible for a musical artifact it must be significantly larger than its expected value, i.e., $A^2(i,k) \gg E[\widehat{A}^2]$. Two methods of determining η are examined. The first method is based on the ratio between the investigated spectral component $\widehat{A}^2(i,k)$ and its expected value $E[\widehat{A}^2]$,

$$\eta_h = \kappa E[\widehat{A}^2], \tag{6.5a}$$

where κ denotes a positive real number. Note that $E[\widehat{A}^2]$ and $\sigma_{\widehat{A}^2}$ are generally unknown statistics, however, they can be estimated based the neighbourhood $\widehat{\mathbf{A}}$. A hypothesis-test which incorporates thresholds η_h will provide a non-intrusive measure, there it does not require the clean speech signal to be known.

The second method of determining threshold η implies the probability of spectral component $\widehat{A}^2(i,k)$ being an outlier, $Pr(\widehat{A}^2(i,k) \in \mathrm{H}_1)$. If we assume the probability of $\widehat{A}^2(i,k)$ to represent an outlier responsible for a musical artifact to be equal to a certain probability q, then the problem of finding a suitable representation for threshold η becomes a problem of finding a suitable representation for q. This threshold, η_s, can be calculated of as follows,

$$\begin{aligned} q &= Pr(\widehat{A}^2(i,k) \in \mathrm{H}_1) \\ &= \int_{\eta_s}^{\infty} p_{\widehat{A}^2}(r^2;\beta_{ss})dr^2, \end{aligned} \tag{6.6a}$$

with $p_{\widehat{A}^2}$ as derived in Eq. 2.14,

$$\begin{aligned} &= \int_{\eta_s}^{\infty} \frac{1}{\sigma_Y^2} \exp\left(-\frac{r^2+\beta_{ss}E[W^2]}{\sigma_Y^2}\right) + \left(1-\exp\left(-\frac{\beta_{ss}E[W^2]}{\sigma_Y^2}\right)\right)\delta[r^2]dr^2 \\ &= \exp\left(-\frac{\eta_s+\beta_{ss}E[W^2]}{\sigma_Y^2}\right) = q, \end{aligned}$$

which leads to,

$$\eta_s = \sigma_Y^2 \ln(1/q) - \beta_{ss}E[W^2]. \tag{6.6b}$$

The physical meaning of parameter q is the relative number of spectral coefficients that are classified as isolated spectral peaks. Consequently, q should dependent on the over-subtraction parameter β_{ss}. To ensure this dependency, parameter q is to be a mapping from the results obtained in the listening experiments.

First, a representation of the number of musical artifacts in a stimulus needs to be found, by using the outcome of the listening experiments as presented in Fig. 3.2(a). There it was argued in Chapter 2.6 that the number of musical artifacts is independent of the noise variance, the average MOMN-score over SNR will provide a suitable representation of the number of musical artifacts. This representation of the number of musical artifacts is depicted in Fig. 6.1(a).

Now, probability $q(\beta_{ss})$ can be shaped using the results depicted in Fig. 6.1(a), by fitting a function $F(\beta_{ss})$ to the data. An exponential function was found to provide a good fit, i.e,

$$F(\beta_{ss}) = a_1 \exp(a_2\beta_{ss}) + a_3 \exp(a_4\beta_{ss}), \tag{6.7}$$

where a_1, a_2, a_3 and a_4 are to be obtained using a non-linear least squares method, $\min_a \|F(\beta_{ss}) - \text{MOMN-score}(\beta_{ss})\|_2^2$. Fig. 6.2 depicts F subject to β_{ss}.

However, there exists an upper limit for $q(\beta_{ss})$, as $q(\beta_{ss})$ can only be equal or smaller than the probability mass in the skirt of the enhanced PDF. Fig. 6.3 depicts the maximum choice for q subject to β_{ss}, moreover, an analytical expression can be found as,

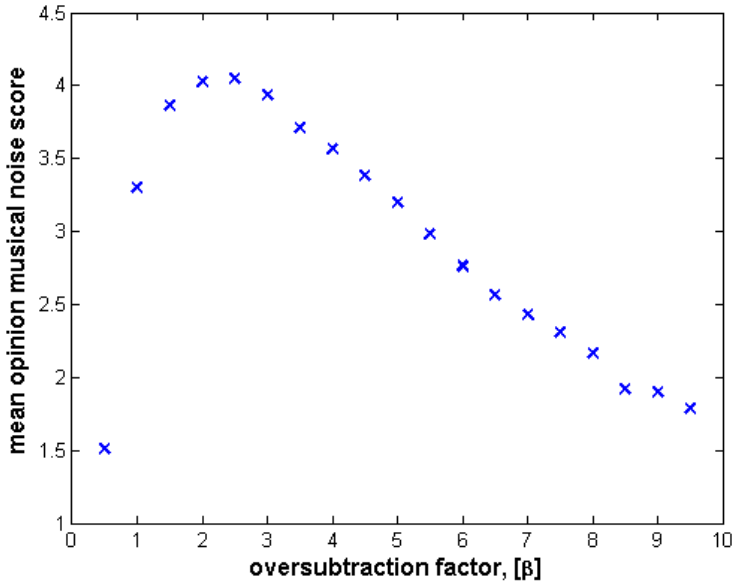


Fig. 6.1: Results of the listening experiments averaged over SNR, providing a representation of the number of musical artifacts in terms of MOMN-score

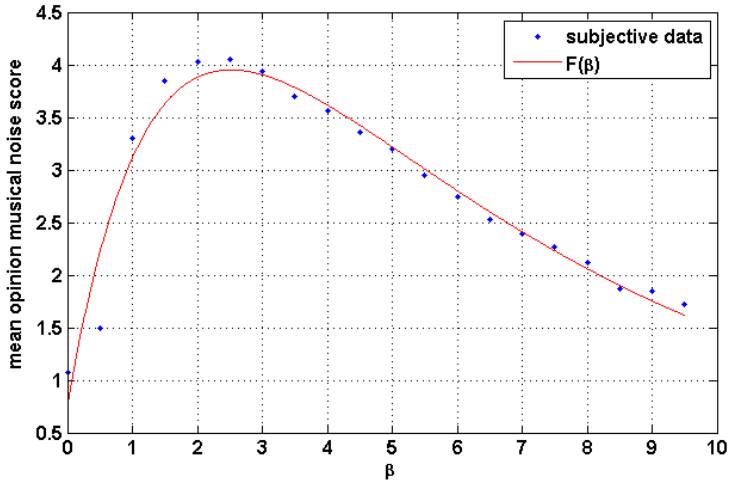


Fig. 6.2: Subjective data and exponential model $F(\beta_{ss})$. $a_1 = 5.63$, $a_2 = -0.13$, $a_3 = -5.3$, $a_4 = -1.33$ (R-square $= 0.96$ [47]).

$$
\begin{aligned}
q_{\max}(\beta_{ss}) &= \lim_{\zeta \downarrow 0} \int_{\zeta}^{\infty} p_{\widehat{A^2}}(r^2; \beta_{ss}) dr^2 \\
&= \lim_{\zeta \downarrow 0} \int_{\zeta}^{\infty} \frac{1}{\sigma_Y^2} \exp\left(-\frac{r^2 + \beta_{ss} E[W^2]}{\sigma_Y^2} \right) + (1 - \exp\left(-\frac{\beta E[W^2]}{\sigma_Y^2} \right)) \delta[r^2] dr^2 \\
&= \lim_{\zeta \downarrow 0} \exp\left(-\frac{\zeta + \beta_{ss} E[W^2]}{\sigma_Y^2} \right) \\
&= \exp\{ -\frac{\beta_{ss} E[W^2]}{\sigma_Y^2} \}.
\end{aligned}
$$

(6.8)

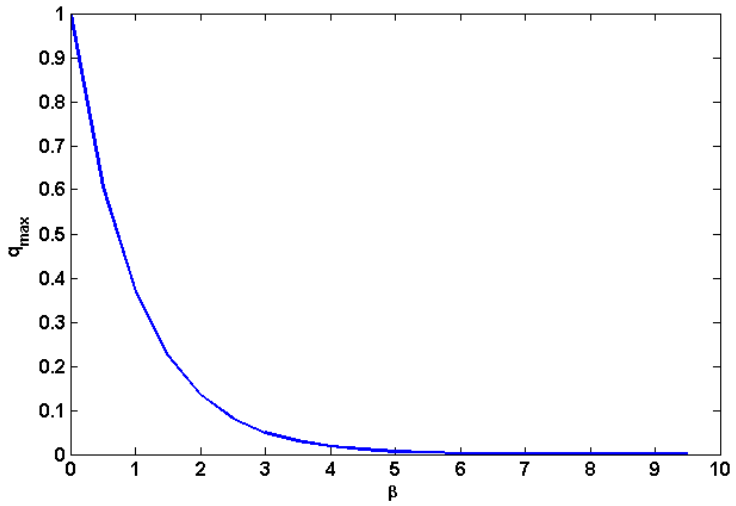


Fig. 6.3: $q_{max}(\beta_{ss})$; $q(\beta_{ss})$ is limited to the probability mass in the skirt of the enhanced PDF.

If $F(\beta_{ss})$ is incorporated to provide a suitable description of $q(\beta_{ss})$, then $q(\beta_{ss})$ should be limited by $q_{\max(\beta_{ss})}$. Further, to provide a representation of $q(\beta_{ss})$, $F(\beta_{ss})$ is to be normalized by Δ(MOMN-score), and scaled using parameter κ_s, that is,

$$\begin{aligned} q(\beta_{ss}) &= \min\left(\frac{\kappa_s F(\beta_{ss})}{\Delta(\text{MOMN-score})}, q_{\max}(\beta)\right) \\ &= \min\left(\frac{\kappa_s F(\beta_{ss})}{\max(\text{MOMN-score}) - \min(\text{MOMN-score})}, q_{\max}(\beta_{ss})\right) \\ &= \min\left(\frac{\kappa_s F(\beta_{ss})}{4}, q_{\max}(\beta_{ss})\right). \end{aligned} \tag{6.9}$$

where, κ_s can be an arbitrarily chosen value between 0 and 1; it was found that $\kappa = 0.01$ is a suitable value. Fig. 6.4 depicts $q(\beta_{ss})$ for $\kappa_s = 0.01$. Now, $q(\beta_{ss})$ denotes the probabiliy of $\widehat{A}^2(i,k)$ being an outlier.

Additionally, it must be noted that in case $q(\beta_{ss}) = q_{\max}(\beta_{ss})$, threshold η_s will be equal to 0,

$$\begin{aligned} \eta_s &= \sigma_Y^2 \ln(1/q_{\max}(\beta_{ss})) - \beta_{ss} E[W^2] \\ &= \sigma_Y^2 \ln\left(\frac{1}{\exp\left(-\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)}\right) - \beta_{ss} E[W^2] \quad = 0. \end{aligned}$$

This indicates that the hypothesis-test which makes use of threshold η_s, would only be applicable if $q_{\max}(\beta_{ss}) > q(\beta_{ss})$. Thus, for certain values of β_{ss} all

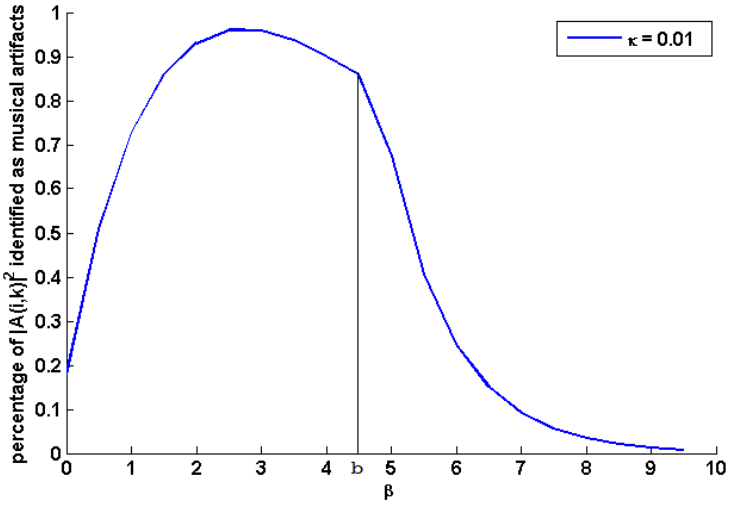


Fig. 6.4: Percentage of musical artifacts detected, β_{ss} ($\kappa_s = 0.01$).

remaining spectral components, including speech components, will be classified as musical artifacts, which is an undesirable property for any musical noise predictor. A musical noise measure which uses threshold η_s, is only be applicable for limited values of β_{ss}.

Consequently, if enhanced signals are given but the exact settings (β_{ss}) of the noise reduction algorithm to these signals are unknown, then this musical noise measure cannot be used. This can however be solved by applying the musical noise measure solely to noise-only signal segments. That is, it can be argued that for a certain value $\beta_{ss} \geq b$ the background noise will consist of musical noise only, if κ_s is calculated such that $q_{\max}(b) = q(b)$, then from $\beta_{ss} > b$ every residual noise component will be detected as an outlier. As a consequence of implying knowledge of the clean speech signal the musical noise measure becomes an intrusive metric. Finding a suitable value for b requires additional listening experiments, and therefore it will remain future work.

7. EVALUATION OF THE OBJECTIVE MEASURES

In this chapter we will describe the evaluation of the performance of the instrumental measures which are considered in Chapter 5 and Chapter 6. The instrumental measures are evaluated using the method described in Chapter 4.

7.1 Evaluation Of The Standard Instrumental Measures

If an instrumental measure is able to predict the outcome of the listening experiment with high correlation, then the metric should, preferably, not contradict its original purpose as a speech-quality or speech-intelligibility measure. That is, an enhanced speech signal with a high measured quality or intelligibility, should contain a minimum amount of musical noise. If this is the case, then the instrumental measure could be used to optimize the parameters of a noise reduction method for both the minimum amount of musical noise and the best speech-quality or speech-intelligibility simultaneously. This is possible if the measured correlation is close to a desired correlation coefficient ρ_d. This desired correlation coefficient is subject to the physical properties of an instrumental measure.

For example, consider a spectral distance measures, where small distances between a clean and an enhanced noisy speech signal indicate good speech-quality. Optimizing SS for a minimum amount of musical noise and minimal spectral distance simultaneously is possible if the spectral distance between the enhanced noisy speech signal and the clean speech signal is large, and if the amount of musical noise in the enhanced noisy speech signal is also large. Hence, a spectral distance measures will be suitable for optimization if $\rho_d \approx 1$. Similar argumentation can be provided for all instrumental measures described in Chapter 5.

Table 7.1 provides the performance of the instrumental measures described in Chapter 5 in terms of their ability to predict the outcome of the listening experiments as presented in Chapter 3, with exception of kurtR which will be discussed in extend in Section 7.2.

It can be observed from Table 7.1, that four instrumental measures correlate strongly with the outcome of the listening experiment, namely, the EUCL, fwSNRn, PESQ and STOI measure. Reviewing ρ provides a surprising result, as the performance of the four instrumental measure indicate that a small spectral distance, high SNR, high speech-quality and high speech-intelligibility is obtained if the stimulus contains a large amount of musical noise. Moreover, this indicates that it is not possible to optimize β_{ss} for both the minimal amount of musical noise and the maximum speech-quality or intelligibility, simultaneously. On the contrary, the results suggest that one should look for the maximal amount of

	input SNR					input SNR			
Measure	**0 dB**	**5 dB**	**10 dB**	**15 dB**		**0 dB**	**5 dB**	**10 dB**	**15 dB**
	ρ	ρ	ρ	ρ	$\rho_d \approx$	τ	τ	τ	τ
EUCL	**-0.79**	**-0.92**	**-0.88**	**-0.81**	1	-0.73	-0.60	-0.58	-0.50
MSD	-0.4	-0.43	-0.33	-0.26	1	-0.23	-0.26	-0.08	-0.25
MSD2	**-0.79**	-0.64	-0.61	-0.52	1	-0.73	-0.60	-0.56	-0.50
LSD	-0.62	-0.63	-0.64	-0.57	1	-0.68	-0.63	-0.61	-0.56
LLR	-0.59	-0.63	-0.64	-0.61	1	-0.44	-0.61	-0.61	-0.66
IS	0.45	0.39	0.37	0.33	1	0.58	0.46	0.48	0.44
SNRseg	0.08	0.15	0.22	0.36	-1	-0.5	-0.33	-0.14	0.11
fwSNR	0.42	0.55	**0.73**	**0.91**	-1	0.33	0.61	0.78	0.94
fwSNRn	0.67	**0.76**	**0.79**	**0.71**	-1	0.69	0.71	0.67	0.64
CEP	-0.29	-0.44	-0.56	-0.57	1	-0.21	-0.41	-0.58	-0.65
PESQ	0.69	**0.91**	**0.96**	**0.97**	-1	0.34	0.85	0.87	0.91
COMPovl	-0.10	0.21	0.45	0.49	-1	-0.11	-0.01	025	0.56
COMPbn	-0.56	-0.48	-0.21	0.16	1	-0.41	-0.38	0.18	0.08
STOI	**0.71**	**0.82**	**0.90**	**0.92**	-1	0.69	0.76	0.80	0.85

Tab. 7.1: Performance of the instrumental measures in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ

musical noise to obtain the best enhanced signal in terms of speech-quality and speech-intelligibility.

7.2 Evaluation Of The Kurtosis Based Musical Noise Measure

The performance of the musical noise metric proposed by Uemura et al. [54] is shown in Table 7.2, where kurtR_1 is derived using the definition of kurtosis denoted in Eq. (5.12) and, kurtR_2 is derived using the definition of kurtosis described in Eq. (5.14). There the kurtosis based metric can only be applied to noise-only coefficients, the VAD as described in Eq. (5.1) is used to exclude speech coefficients in the enhanced noisy-speech signal.

Uemura et al. [54] investigated the correlation between the outcome of a listening experiment and the kurtosis ratio as described in Eq (5.11). A logarithmic mapping of kurtR was used in [54], i.e ln(kurtR), to derive the performance of the metric for $0 < \beta_{ss} < 2$. This mapping corrects for a non-linear relationship between the subjective data and kurtR.

Considering this logaritmic mapping and the range $0 < \beta_{ss} < 2.5$ will provide an identical experiment as was performed in [54]. It can be concluded from Table 7.2 that within the rance $0 < \beta_{ss} < 2.5$, kurtR will provide a strong correlation with the outcome of the listening experiment in Chapter 3, which is consistent with the observation made in [54]. However, for $\beta_{ss} > 2.5$, kurtR will keep increasing, in contradiction to the results obtained from the listening experiment, which show a decrease in the amount of musical noise. This causes the metric to perform with low correlation for $\beta_{ss} > 2.5$. Overall, if β_{ss} cannot be guaranteed, e.g., if the processing algorithm is unknown, the KBMN metric may perform poorly.

Considering a noise only segment, it can be seen that kurtosis as defined in Eq. (5.12) and Eq. (5.13) will be insensitive to scaling of the noise process, in contradiction to kurtosis as defined in Eq. (5.14). That is, the kurtosis determined

	input SNR							
Measure	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
	ρ	ρ	ρ	ρ	τ	τ	τ	τ
	$0 < \beta_{ss} < 2.5$							
kurtR_1	**0.71**	**0.72**	**0.74**	**0.89**	0.86	0.86	0.73	1
$\ln(\text{kurtR}_1)$	**0.89**	**0.89**	**0.90**	**0.97**	”	”	”	”
kurtR_2	**0.68**	**0.69**	**0.71**	**0.87**	”	”	”	”
$\ln(\text{kurtR}_2)$	**0.80**	**0.82**	**0.83**	**0.94**	”	”	”	”
	$\beta_{ss} > 2.5$							
kurtR_1	-0.27	-0.29	-0.29	-0.29	-0.31	0.25	0.27	0.20
$\ln(\text{kurtR}_1)$	0.09	0.07	0.06	0.10	”	”	”	”
kurtR_2	-0.27	-0.29	-0.29	-0.29	”	”	”	”
$\ln(\text{kurtR}_2)$	0.08	0.07	0.06	0.09	”	”	”	”
	$0 < \beta_{ss} < 9.5$							
kurtR_1	-0.45	-0.5	-0.53	-0.48	-0.49	-0.46	-0.48	-0.43
$\ln(\text{kurtR}_1)$	-0.34	-0.37	-0.39	-0.34	”	”	”	”
kurtR_2	-0.45	-0.50	-0.53	-0.48	”	”	”	”
$\ln(\text{kurtR}_2)$	-0.41	-0.43	-0.45	-0.39	”	”	”	”

Tab. 7.2: Performance of the instrumental measures in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

using Eq. (5.12) provided a signal $X_2 \sim \mathcal{CN}(0, c^2\sigma^2_{X_1})$ transformed into the PDFT domain,

$$\begin{aligned}\mathcal{K}_{X_2^2} &= \frac{\mu_4}{\mu_2^2} \\ &= \frac{E[(X_2^2)^4]}{\left(E[(X_2^2)^2]\right)^2} \\ &= \frac{c^8 E[(X_1^2)^4]}{c^8\left(E[(X_1^2)^2]\right)^2} \\ &= \frac{E[(X_1^2)^4]}{\left(E[(X_1^2)^2]\right)^2} \\ &= \mathcal{K}_{X_1^2}.\end{aligned}$$

Similarly, $\mathcal{K}_{X_2^2}$ equals $\mathcal{K}_{X_1^2}$ when calculated using Eq. (5.13), however, $\mathcal{K}_{X_2}$ determined using Eq. (5.14) will result in $\mathcal{K}_{X_2} = \mathcal{K}_{X_1} - 2$.

In Chapter 2.6 and 3 it was concluded that the amount of musical noise is sensitive to scaling of the noise process. Upon evaluating Eq. (5.12) and Eq. (5.13) in the scope of the definition of the amount of musical noise as provided in Chapter 2.6, it can thus be argued that, due to the invariance to scaling of the noise process, both methods can not be incorporated to derive kurtR.

If the number of isolated spectral peaks could be predicted by means of the metric kurtR and, if kurtR is derived by implying the kurtosis as denoted Eq. (5.12), then it can be argued that for certain values of β_{ss}, kurtR will not provide a suitable prediction of the number of musical artifacts. In [54], X was assumed to represent a noise-only signal in the power spectral domain, consequently, $E[W^2] = \sigma_Y^2$, where $\sigma_Y^2 = \sigma^2_{\text{Re}(V)} + \sigma^2_{\text{Im}(V)}$. Provided a generalized

description of the PDF of signal X^2 (see Eq. (2.14)) a description of kurtosis can be easily found,

$$p_{\widehat{X}^2}(x^2) = \begin{cases} \frac{1}{\sigma_Y^2} \exp\left(-\frac{x^2+\beta E[W^2]}{\sigma_Y^2}\right) + \left(1 - \exp\left(-\frac{\beta E[W^2]}{\sigma_Y^2}\right)\right) \delta[x^2] & x^2 \geq 0 \\ 0 & otherwise \end{cases} \tag{7.1}$$

If $\beta_{ss} = 0$ then Eq. (7.1) describes the PDF of the noisy speech coefficients R^2, further, if $\beta_{ss} > 0$ then Eq. (7.1) provides the PDF of the enhanced noise speech coefficients $\widehat{A}^2$.

First, μ_4 and μ_2 can be determined as,

$$\mu_4 = \int_0^\infty t^4 p_{\widehat{X}^2}(t)dt = 24[\sigma_Y^2]^4 \exp\left(-\frac{\beta E[W^2]}{\sigma_Y^2}\right),$$
$$\mu_2 = \int_0^\infty t^2 p_{\widehat{X}^2}(t)dt = 2[\sigma_Y^2]^2 \exp\left(-\frac{\beta E[W^2]}{\sigma_Y^2}\right),$$

where $t = x^2$, after which the kurtosis of the PDF of X can be found by,

$$\mathcal{K}_X = \frac{24[\sigma_Y^2]^4 \exp\left(-\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)}{\left(2[\sigma_Y^2]^2 \exp\left(-\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)\right)^2} = 6 \exp\left(\frac{\beta E[W^2]}{\sigma_Y^2}\right). \tag{7.2}$$

Consequently, for $\beta_{ss} = 0$ the kurtosis is equivalent to $\mathcal{K}_{R^2} = 6$, and for $\beta_{ss} > 0$ the kurtosis of the enhanced signal becomes equivalent to $\mathcal{K}_{\widehat{A}^2} = 6 \exp(\beta_{ss})$. Then, by dividing the kurtosis of the enhanced speech process by the kurtosis of the noisy speech process kurtR can be found,

$$\text{kurtR} = \exp(\beta_{ss}). \tag{7.3}$$

In this case kurtR represents an exponentially increasing function subject to the over-subtraction parameter β_{ss}.

An identical result can be found in case the kurtosis is estimated using Eq. (5.14). That is, given that $E[X^4] = E[X^2]$, $\mathcal{K}_X$ can be casulated as follows,

$$\mathcal{K}_X = \frac{[\sigma_Y^2] \exp\left(-\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)}{[\sigma_Y^2] \exp\left(-\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)^2} = [\sigma_Y^2]^{-1} \exp\left(\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right).$$

Now, kurtR becomes equal to Eq. (7.3),

$$\text{kurtR} = \frac{[\sigma_Y^2]^{-1} \exp\left(\frac{\beta_{ss} E[W^2]}{\sigma_Y^2}\right)}{[\sigma_Y^2]^{-1}} = \exp(\beta_{ss}). \tag{7.4}$$

The case of estimating kurtR by using the definition of kurtosis as defined in Eq. (5.13) provides a somewhat different result, however, the same conclusion can be

drawn as for the case in which kurtR is calculated by implying the definition of kurtosis given in Eq. (5.12) or Eq. (5.14). Here,

$$E\left[(X^2 - E[X^2])^4\right] = 24[\sigma_Y^2]^4 \exp\left(-\frac{\beta_{ss}E[W^2] + E[X^2]}{\sigma_Y^2}\right),$$

and,

$$\left(E\left[(X^2 - E[X^2])^2\right]\right)^2 = 4[\sigma_Y^2]^4 \exp\left(-\frac{2(\beta_{ss}E[W^2] + E[X^2])}{\sigma_Y^2}\right).$$

Then kurtR can be denoted as,

$$\text{kurtR} = \exp\left(\beta_{ss} + [\sigma_Y^2]^{-1}E[X^2]\right) = \exp\left(\beta_{ss} + \exp(-\beta_{ss})\right). \tag{7.5}$$

which is again exponentially increasing function of the over-subtraction parameter β_{ss}.

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 kurtR becomes undefined. For example, consider a situation in which the noisy speech signal X is enhanced using SS and, β_{ss} is chosen equal to infinity, then no signal would be audible whatsoever. The musical noise measure as proposed by Uemura et al. [54] will in fact predict an infinite amount of musical noise. Hence, it can be argued that kurtR, if defined as a measure of the number of musical artifacts, is not applicable for all values of β_{ss} as was concluded from Table 7.2. Fig 7.1 depicts the two methods of calculating kurtR for various values of β_{ss}.

Consequently, kurtR can only be applicable as a musical noise measure if the kurtosis is estimated using the definition provided in Eq. (5.14), moreover, kurtR will only provide a valid musical noise measure for certain values of β_{ss}.

7.3 Evaluation Of The Outlier Based Musical Noise Measure

All results have been obtained using a 5×5 sized window, $\widehat{\mathbf{A}}^2$, centered around $\widehat{A}^2(i,k)$, describing the neighbourhood around the enhanced spectral component $\widehat{A}^2(i,k)$.

OBMN measure applying η_h

When applying the hypothesis-test based method which implies η_h to the signals in the experiment data-set $\{\widehat{s}_{\text{SNR},\ \beta_{ss}}[n]\}$ provided in Eq. (3.1), then the parameter κ is to be chosen arbitrarily. However, κ is to be chosen such, that the number of musical artifacts (Fig 6.1(a)) is detected with high correlation. From Fig. 7.2 it can be observed that $\kappa > 7$ will produce a well performing metric.

This bound, $\kappa > 7$, can be used to calculate the amount of musical noise, following the procedure described in Eq. 6.4. Table 7.3 depicts the performance of the proposed musical noise predictor for the various values of κ.

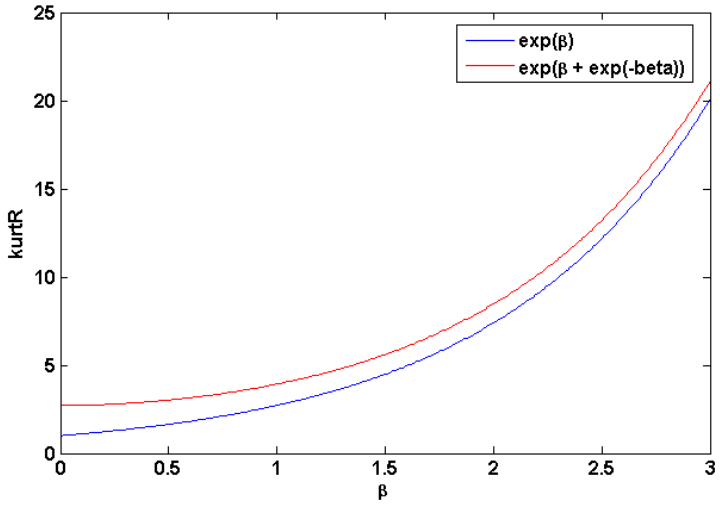


Fig. 7.1: The KBMN measure, kurtR, for different interpretations of the kurtosis

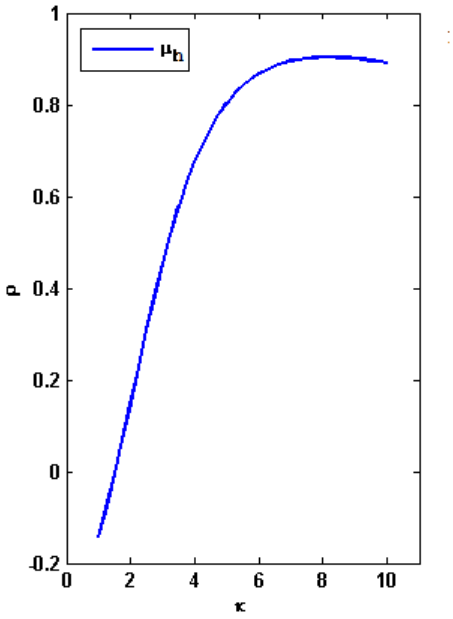


Fig. 7.2: Performance in terms of prediction of the number of musical artifacts (Fig 6.1(a)), for various values of κ.

There exists the possibility that a speech component is falsely classified as an isolated spectral peak. Consider, for example, an enhanced noisy speech signal $\widehat{A}^2(i,k) = R^2(i,k) - \beta_{ss} E[W^2]$. If $R^2(i,k)$ has got a high SNR, then the energy of the falsely classified components will have large influence on the predicted amount

	input SNR							
	0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
κ	ρ	ρ	ρ	ρ	τ	τ	τ	τ
7	**0.88**	0.69	0.23	0.21	0.91	0.28	0.29	-0.40
7.5	**0.91**	0.63	0.40	-0.02	0.93	0.46	-0.35	-0.35
8	**0.85**	**0.77**	0.63	0.31	0.78	0.29	0.47	-0.15
8.5	**0.86**	**0.81**	0.63	0.40	0.71	0.63	0.48	-0.01
9	**0.88**	**0.83**	0.57	0.14	0.89	0.55	0.32	-0.15
9.5	**0.88**	**0.85**	0.53	0.10	0.91	0.73	0.33	-0.36

Tab. 7.3: performance of the OBMN measure incorporating threshold η_h in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

of musical noise. This explains why, for signals having high input SNR, i.e. 10 dB and 15 dB, the outcome of the musical noise metric shows low correlation with the listening test data.

This latter conclusion can be supported by investigating whether the musical noise predictor will provide better performance, if solely the noise-only coefficients of the enhanced noisy-speech is used to estimate the amount of musical noise. From the results denoted in Table 7.4 which shows the performance of the OBMN measure which incorporates threshold η_h, it can be seen that the metric provides better performance for high SNR if the clean speech process is known. The noisy-only PDFT coefficients are found by applying the VAD described in Eq. (5.1). In both these cases τ supports the conclusions made.

	input SNR							
	0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
κ	ρ	ρ	ρ	ρ	τ	τ	τ	τ
7	0.59	0.59	0.58	0.53	0.63	0.60	0.62	0.57
7.5	0.63	0.62	0.63	0.58	0.65	0.62	0.64	0.59
8	0.66	0.66	0.67	0.62	0.65	0.62	0.64	0.60
8.5	0.69	0.69	**0.70**	0.66	0.68	0.65	0.67	0.62
9	**0.72**	**0.72**	**0.74**	**0.72**	0.68	0.65	0.66	0.63
9.5	**0.75**	**0.74**	**0.75**	**0.73**	0.70	0.67	0.68	0.66

Tab. 7.4: Performance of the OBMN measure incorporating threshold η_h in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

OBMN measure applying η_s

The performance of the musical noise measure which implies threshold η_s can be investigated in a similar fashion. It has been argued in Chapter 6 that this method can only be applied on noise-only time-frequency coefficients. To examine the performance of this measure, the VAD as described by Eq. (5.1) is applied to obtain the enhanced noise-only PDFT coefficients.

	input SNR							
	0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB** (τ)	**5 dB**	**10 dB**	**15 dB**
κ	ρ	ρ	ρ	ρ	τ	τ	τ	τ
0.05	0.36	0.35	0.36	0.31	0.58	0.56	0.58	0.53
0.01	0.61	0.61	0.62	0.59	0.65	0.63	0.67	0.62
0.005	0.67	0.68	0.69	0.67	0.71	0.68	0.71	0.66
0.001	**0.77**	**0.78**	**0.79**	**0.82**	0.81	0.78	0.78	0.78

Tab. 7.5: Performance of the OBMN measure incorporating threshold η_s in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

Table 7.5 reveals that the musical noise measure, where $\kappa = 0.005$, performs insufficiently. However, choosing $\kappa = 0.001$ will provide a working musical noise predictor.

7.3.1 Non-linear mapping of the OBMN measure

To produce the results described in Table 7.5 we assume a linear mapping between probability $q(\beta)$ and the perceived number of musical artifacts, however, a non-linear mapping could be more suitable, e.g., $q(\beta) \sim 10^{\mathrm{MOMNS}}$. Upon investigating if a better performance can be be obtained if such a non-linear relationship between the subjective data en the musical noise metric is assumed, we observe the performance of both methods if a logarithmic mapping is applied to the outcome of the predictor, i.e., $\log_{10}$(amount of musical noise). We can think of this mapping as deriving the energy of the detected isolated spectral peaks in [dB/10]. Again the two methods of finding threshold η are investigated. The performance results are shown in Table 7.6 and 7.7. In the case threshold η_s is implied, no performance increase can be observed, however, for the metric which makes us of η_s, we find that for $\kappa \leq 0.01$ a well performing musical noise predictor can be obtained.

	input SNR							
	0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
κ	ρ	ρ	ρ	ρ	τ	τ	τ	τ
7.5	0.65	0.67	0.68	0.63	0.65	0.62	0.64	0.59
8	0.67	0.69	**0.70**	0.65	0.65	0.62	0.64	0.60
8.5	**0.70**	**0.71**	**0.72**	0.68	0.68	0.65	0.67	0.62
9	**0.70**	**0.72**	**0.74**	**0.70**	0.68	0.65	0.66	0.63
9.5	**0.72**	**0.74**	**0.75**	**0.71**	0.70	0.67	0.68	0.66

Tab. 7.6: Performance of the logarithmic mapping of the OBMN measure incorporating threshold η_h in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

	input SNR							
	0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB** (τ)	**5 dB**	**10 dB**	**15 dB**
κ	ρ	ρ	ρ	ρ	τ	τ	τ	τ
0.05	0.56	0.58	0.60	0.56	0.58	0.56	0.58	0.53
0.01	**0.71**	**0.71**	**0.74**	**0.70**	0.65	0.63	0.67	0.62
0.005	**0.77**	**0.78**	**0.80**	**0.76**	0.71	0.68	0.71	0.66
0.001	**0.88**	**0.89**	**0.90**	**0.88**	0.81	00.78	0.78	0.78

Tab. 7.7: Performance of the logarithmic mapping of the OBMN measure incorporating threshold η_s in terms of the Pearson correlation coefficient ρ and Kendall's tau correlation parameter τ.

8. CONCLUSIONS AND FUTURE WORK

First, 80 different enhanced noisy speech signals were obtained using the spectral subtraction noise reduction method, all these signals were created in such a way that each signal contained a different amount of musical noise. In order to obtain a ground truth on the amount of musical noise present in these signals, we conducted a listening experiment which was especially designed to quantify the amount of musical noise in stimuli.

Various instrumental measures have been evaluated on their ability to predict the outcome of the listening experiment. Upon evaluating the performance of these instrumental measures, it was shown that four different standard instrumental measures perform well as musical noise measures, i.e, EUCL, fwSNRn, PESQ and STOI. Surprisingly, the results suggested that one should optimize the over-subtraction factor for the maximum amount of musical noise to obtain the optimal enhanced signal in terms of speech-quality or speech-intelligibility. Furthermore, it was concluded that the KBMN measure proposed in [54], which was specifically designed to predict the amount of musical noise in a stimulus, performs poorly for larger values of β_{ss}. These conclusions were supported by an analysis of the KBMN measure. The applicability of the KBMN measure is therefore very limited, as one requires information about how the enhanced noisy speech signal is produced; β_{ss} must be a known at the predictor.

To gain more insight into of the amount of musical noise in a stimulus we proposed a metric which specifically targets the musical noise. This measure is based on the definition of the amount of musical noise and interprets an isolated spectral peak as an outlier of the probability distribution of an enhanced noisy speech signal. The measure represents a parametric outlier detection method, and as a consequence, a certain threshold η is to be chosen. We proposed two methods of obtaining this threshold, which resulted in an intrusive musical noise measure incorporating threshold η_h, and a non-intrusive musical noise measure which incorporates threshold η_s. An extensive description is provided on how these thresholds were obtained, however, in the case of η_s, we do not provide the optimal way of finding the mapping of the listening-test data to the threshold, nor finding the optimal value for κ. Finding the optimal threshold η will remain future work. Furthermore, note that the OBMN measure will incorporate all classified spectral components $\widehat{A}^2(i,k)$, as audible musical artifacts. However, as stated in Chapter 2.6, not all these isolated spectral components will be perceptible due to the masking characteristics of the human auditory system. This could be solved by considering the auditory masking threshold, but is beyond the scope of this thesis.

Both proposed implementations of the OBMN measure provide well performing musical noise measures, there correlations of $0.7 \leq \rho \leq 0.9$ can be obtained

for certain configurations. Additionally, we found that a non-linear relationship between the OBMN measures and the outcome of the listening experiment provides the best performance, i.e., $\rho \approx 0.9$ for the OBMN measure implying η_s. It must be noted that, for signals with high input SNR, misclassified speech components have strong influence on the performance of the non-intrusive measure (OBMN measure which implies threshold η_h). Results obtained when an VAD based on the clean speech signal was implied, making this non-intrusive measure an intrusive measure, showed increased performance for high input SNR.

Comparing the overall performance of the standard instrumental measures, the KBMN measure and the OBMN measure, it can be concluded that the OBMN measure which incorporates threshold η_s provides, overall, the best performing musical noise measure, as it is the instrumental measure which is able to predict the outcome of the listening experiment with highest correlation .

This research could be extended by taking into account a selection of frequently used noise reduction methods, with the purpose to investigate if a maximal amount of musical noise always occurs for high quality enhanced signals. One could then draw conclusions on the instrumental measures in terms of their sensitivity to musical noise. In practice, the proposed OBMN measure could be used along side an established instrumental measures to provide better predictions of the speech-quality or speech-intelligibility. For this latter to be possible, the performance of the OBMN measure is to be investigated for noisy signals enhanced with various different noise reductions methods, furthermore, it needs to be shown that such a combination of instrumental measures will provide a better prediction of the speech-quality or speech-intelligibility.

APPENDIX

A. ANALYSIS OF THRESHOLD η

The proposed musical noise predictor estimates the perceived amount of musical noise by incorporating an estimate of the number of musical artifacts. When predicting the number of musical artifacts the measure should be insensitive to scaling of the noise process. To analyze this property we will consider the following examples.

If $y[n]$ denotes a noise-only signal, $y[n] = cv[n]$, the detection problems incorporating η_h, described in Eq. (6.5a), will be linearly related to scaling of the noise process components by means of c.

Example
Consider a scaled zero mean complex gaussian distributed process in the DFT domain, $Y_2(i,k) = V_2(i,k) = cV_1(i,k)$, *i.e.* $Y_2(i,k) \sim \mathcal{CN}(0, \sigma^2_{V_2}) = \mathcal{CN}(0, c^2\sigma^2_{V_1})$. $Y(i,k)$ *is then transformed into the PDFT domain and processed using SS. The underlying PDF of* $\widehat{A}^2(i,k)$ *is denoted by Eq.* (2.14). *As was argued in Chapter 2.6, the enhanced spectral component* $\widehat{A}_2^2(i,k)$ *will be a linearly scaled version of* $\widehat{A}_1^2(i,k)$,

$$\widehat{A}_2^2(i,k) = c^2 \widehat{A}_1^2(i,k). \tag{A.1}$$

As a consequence of the assumptions made above $E[R^2] = E[W^2] = \sigma_Y^2$, *where* $\sigma_Y^2 = \sigma^2_{Re(V)} + \sigma^2_{Im(V)}$, *and thus the expected value of* $\widehat{A}_2^2(i,k)$ *can be denoted as a scaled version of* $E[\widehat{A}_1^2]$, *i.e.,*

$$\begin{aligned} E[\widehat{A}_2^2] &= \sigma^2_{Y_2} \exp\left(-\frac{\beta E[{W_2}^2]}{\sigma^2_{Y_2}}\right) \\ &= \sigma^2_{Y_2} \exp\left(-\frac{\beta \sigma^2_{Y_2}}{\sigma^2_{Y_2}}\right) \\ &= \sigma^2_{Y_2} \exp\left(-\beta\right) \\ &= c^2 \sigma^2_{Y_1} \exp\left(-\beta\right) \\ &= c^2 E[\widehat{A}_1^2]. \end{aligned} \tag{A.2}$$

Likewise, also $\sigma_{\widehat{A}_2^2}$ *is linearly related to* $\sigma_{\widehat{A}_1^2}$*, i.e.,*

$$\begin{aligned}
\sigma_{\widehat{A}_2^2} &= \sqrt{\left[\sigma_{Y_2}^2\right]^2 \left(2\exp\left(-\frac{\beta E[{W_2}^2]}{\sigma_{Y_2}^2}\right) - \exp\left(-\frac{2\beta E[{W_2}^2]}{\sigma_{Y_2}^2}\right)\right)} \\
&= \sqrt{\left[\sigma_{Y_2}^2\right]^2 \left(2\exp\left(-\frac{\beta\sigma_{Y_2}^2}{\sigma_{R_2}^2}\right) - \exp\left(-\frac{2\beta\sigma_{Y_2}^2}{\sigma_{Y_2}^2}\right)\right)} \\
&= \sqrt{\left[\sigma_{Y_2}^2\right]^2 \left(2\exp(-\beta) - \exp(-2\beta)\right)} \\
&= c^2\sqrt{\left[\sigma_{Y_1}^2\right]^2 \left(2\exp(-\beta) - \exp(-2\beta)\right)} \\
&= c^2\sigma_{\widehat{A}_1^2}.
\end{aligned} \tag{A.3}$$

By incorporating the conclusions made in Eq. (A.1)*,* (A.2) *and.* (A.3)*, it can concluded that Eq.* (6.5a) *is linearly related to* c^2*. Thus,* $H_1 : \widehat{A}_2^2(i,k) > \eta$ *is equivalent to,*

$$H_{1,\eta_h} : c^2\widehat{A}_1^2(i,k) > c^2\kappa E[\widehat{A}_1^2], \tag{A.4a}$$

(A.4b)

which shows that the outlier detection method is insensitive to scaling of the noise process.

Similarly it can be observed, that if $y[n] = cv[n]$, the method incorporating η_s, described by Eq. (6.6b), is insensitive of scaling of the noise source by means of c.

Example

Consider the noise-only DFT component $Y_2(i,k) = V_2(i,k) = cV_1(i,k)$*. Here the underlying PDF of* $Y(i,k)$ *can be described by a scaled zero mean complex Gaussian distribution, i.e.* $Y_2(i,k) \sim \mathcal{CN}(0,\sigma_{Y_2}^2) = \mathcal{CN}(0,c^2\sigma_{Y_1}^2)$*.* $Y(i,k)$ *is transformed into the PDFT domain and processed using SS, the underlying PDF of* $\widehat{A}^2(i,k)$ *is denoted by Eq.* (2.14)*. First, it can be shown that if* $R^2(i,k) = W^2(i,k)$ *and* $E[R^2] = E[W^2] = \sigma_Y^2$*, where* $\sigma_Y^2 = \sigma_{Re(V)}^2 + \sigma_{Im(V)}^2$*, then* $q_{max}(\beta)$ *is independent of* c*,*

$$\begin{aligned}
q_{2,max}(\beta) &= \exp\left(-\frac{\beta E[W_2^2]}{\sigma_{Y_2}^2}\right) \\
&= \exp\left(-\beta\right).
\end{aligned}$$

Since $F(\beta)$ *is independent of scaling,* $q(\beta)$ *will be independent of scaling. Subsequently it can be seen that threshold* $\eta_{s,\widehat{A}^2}$ *is linearly related to* c^2*,*

$$\begin{aligned}
\eta_{s,\widehat{A}_2^2} &= \sigma_{Y_2}^2\ln\left(1/q(\beta)\right) - \beta E[W_2^2] \\
&= c^2\sigma_{Y_1}^2\ln\left(1/q(\beta)\right) - c^2\beta\sigma_{Y_1} \\
&= c^2(\sigma_{Y_1}^2\ln\left(1/q(\beta)\right) - \beta\sigma_{Y_1}). \\
&= c^2\eta_{s,\widehat{A}_1^2}
\end{aligned} \tag{A.5}$$

It was argued that $\widehat{A}_2^2(i,k) = c^2\widehat{A}_1^2(i,k)$*, hence, if* $H_1 : \widehat{A}_2^2(i,k) > \eta_{s,\widehat{A}_1^2}$ *is equivalent to,* $H_1 : c^2\widehat{A}_1^2(i,k) > c^2\eta_{s,\widehat{A}_1^2}$*, from which it can be concluded that the outlier detection method is insensitive to scaling of the noise variance.*

BIBLIOGRAPHY

[1] Voicebox: Speech processing toolbox for matlab. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. Accessed: May 2013.

[2] J. Antoni. The spectral kurtosis: a useful tool for characterising non-stationary signals. *Mechanical Systems and Signal Processing*, 20(2):282 – 307, 2006.

[3] Vic Barnett and Toby Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[4] J. G. Beerends. Extending p.862 PESQ for assessing speech intelligibility. *White contribution COM 12-C2 to ITU-T Study Group 12*, October 2004.

[5] Irad Ben-Gal. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*, pages 131–146. Springer, 2005.

[6] S. Ben Jebara. A perceptual approach to reduce musical noise phenomenon with wiener denoising technique. In *Acoustics, Speech and Signal Processing. IEEE International Conference on ICASSP '06*, volume 3, 2006.

[7] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on '79.*, volume 4, pages 208–211, 1979.

[8] 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.

[9] 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.

[10] A. Davis, S.Nordholm, S.Y. Low, and R. Togneri. A multi-decision sub-band voice activity detector. In EURASIP, editor, *EUSIPCO*, pages 4214–4217, 2006.

[11] L. T. DeCarlo. On the meaning and use of kurtosis. *Psychological Methods*, 2(3):292–307, March 1997.

[12] Nima Derakhshan, Mohsen Rahmani, Ahmad Akbari, and Ahmad Ayatollahi. An objective measure for the musical noise assessment in noise reduction systems. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4429–4432. IEEE, 2009.

[13] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[14] Y. Ephraim and H. L. van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 3(4):251–266, July 1995.

[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.

[16] M. R. J. Gerrits, R. C. Hendriks, N. D. Gaubitch, J. Jensen, and M. S. Pedersen. Evaluation of instrumental measures for the prediction of musical noise in enhanced noisy speech. In *Wic*, 2013.

[17] Douglas M Hawkins. *Identification of outliers*, volume 11. Chapman and Hall London, 1980.

[18] R. C. Hendriks. Probability density functions for speech enhancement. Technical report ICT-2004-05, Delft University of Technology.

[19] 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.

[20] Victoria J Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[21] 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.

[22] Y. Huajun and T. Fingscheidt. Black box measurement of musical tones produced by noise reduction systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4573–4576, 2012.

[23] T. Inoue, H. Saruwatari, K. Shikano, and K. Kondo. Theoretical analysis of musical noise in wiener filtering family via higher-order statistics. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5076–5079, 2011.

[24] Radiocommunication Sector International Telecomminication Uninion. Recommendation bs. 1116-1. *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, 1997.

[25] Telecomminication Sector International Telecomminication Uninion. Recommendation p.562-3. *Subjective performance assessment of telephone band and widehnad digital codecs*, 1998.

[26] Telecomminication Sector International Telecomminication Uninion. Recommendation p.830. *Subjective performance assessment of telephone band and widehnad digital codecs*, 1998.

[27] Telecomminication Sector International Telecomminication Uninion. Recommendation bs 1534-1. *Subjective performance assessment of telephone band and widehnad digital codecs*, 2003.

[28] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. In *Electron. Commun. Jpn*, volume 53, pages 36–43, 1982.

[29] ITUT. Itu-t recommendation p.830. *Subjective performance assessment of telephone band and widehnad digital codecs*, 1996.

[30] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 2003.

[31] J. Jensen and R. Heusdens. Improved subspace based single-channel speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, 15(3):862–871, 2007.

[32] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

[33] A. H. Gray. Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(5):380–391, October 1976.

[34] C. Li and W. Liu. A novel multi-band spectral subtraction method based on phase modification and magnitude compensation. In *Acoustics, Speech and Signal Processing, 2011 IEEE International Conference on*, pages 4760–4763. IEEE, 2011.

[35] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. of the IEEE*, 67(12):1586–1604, December 1979.

[36] Klaus Linhard and Heinz Klemm. Noise reduction with spectral subtraction and median filtering for suppression of musical tones. In *Robust Speech Recognition for Unknown Communication Channels*, 1997.

[37] P. Loizou. *Speech enhancement theory and practice*. CRC Press, 2007.

[38] 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. America*, (125):3387–3405, 2009.

[39] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[40] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(2):137–145, April 1980.

[41] J. Ortega-García and J. González-Rodríguez. Overview of speech enhancement techniques for automatic speaker recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 929–932. IEEE, 1996.

[42] 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.

[43] T. Painter and A. Spanias. Perceptual coding of digital audio. In *Proceedings of the IEEE*, pages 451–513, 2000.

[44] S. R. Quackenbush, T.P Barnwell, and M.A Clements. Objective measures of speech quality.

[45] H. Saruwatari, Y. Ishikawa, Y. Takahash, T. Inoue, K. Shikano, and K. Kondo. Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher order statistics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6):1457–1466, 2011.

[46] H. Saruwatari, S. Kanehara, R. Miyazaki, K.Shikona, and K. Kondo. Musical noise analysis for bayesian minimum mean-square error speech amplitude estimators based on higher-order statistics. In *INTERSPEECH*, 2013.

[47] D. J. Sheskin. *Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, 3rd edition edition, 2004.

[48] 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.

[49] 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.

[50] 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.

[51] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo. Theoretical musical-noise analysis and its generalization for methods of integrating beamforming and spectral subtraction based on higher-order statistics. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 93–96, 2010.

[52] Y. Takahashi, Y. Uemura, H. Saruwatari, K. Shikano, and K. Kondo. Musical noise analysis based on higher order statistics for microphone array and nonlinear signal processing. In *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, pages 229–232, 2009.

[53] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere. A study of complexity and quality of speech waveform coders. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,*, volume 3, page 586 590. IEEE, 1978.

[54] 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.

[55] V. D. Vrabie, P. Granjon, and C. Serviere. Spectral kurtosis: From definition to application. *NSIP*, 20 Mar 2003.

[56] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *Audio and Electroacoustics, IEEE Transactions on*, 15(2):70–73, 1967.

[57] G. Whipple. Low residual noise speech enhancement utilizing time-frequency filtering. In *Acoustics, Speech and Signal Processing (ICASSP), 1994 IEEE International Conference on*, volume 4, pages 1–5/1–8, 1994.

[58] P. C. Yong, S. Nordholm, and H. H. Dam. Trade-off evaluation for speech enhancement algorithms with respect to the a priori snr estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4657–4660, 2012.

[59] K.-C. Tan Z.Goh and B.T.G. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Trans. Speech Audio Processing*, 6(3):287–292, May 1998.