Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 358–376

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Optimization and evaluation of sigmoid function with *a priori* SNR estimate for real-time speech enhancement

Pei Chee Yong*, Sven Nordholm, Hai Huyen Dam

Curtin University, Kent Street, Bentley, WA 6102, Australia

Received 24 April 2012; received in revised form 3 September 2012; accepted 17 September 2012
Available online 27 September 2012

Abstract

In this paper, an *a priori* signal-to-noise ratio (SNR) estimator with a modified sigmoid gain function is proposed for real-time speech enhancement. The proposed sigmoid gain function has three parameters, which can be optimized such that they match conventional gain functions. In addition, the joint temporal dynamics between the SNR estimate and the spectral gain function is investigated to improve the performance of the speech enhancement scheme. As the widely-used decision-directed (DD) *a priori* SNR estimate has a well-known one-frame delay that leads to the degradation of speech quality, a modified *a priori* SNR estimator is proposed for the DD approach to overcome this delay. Evaluations are performed by utilizing the objective evaluation metric that measures the trade-off between the noise reduction, the speech distortion and the musical noise in the enhanced signal. The results are compared using the PESQ and the SNRseg measures as well as subjective listening tests. Simulation results show that the proposed gain function, which can flexibly model exponential distributions, is a potential alternative speech enhancement gain function.

Keywords: Speech enhancement; SNR estimation; Decision-directed approach; Sigmoid function; Objective evaluation

1. Introduction

Assistive Listening Devices are in high demand in practice for scenarios with high noise such as in loud industrial factories. These scenarios are characterized by fast changing noise in volume and character, with technology devices requiring low power real-time implementation and low processing delay. The noisy speech often results in lower intelligibility and listener fatigue. As such, it is important to develop an efficient speech enhancement algorithm that reduces additive noise from noisy speech while preserving speech components as undistorted as possible. In this paper we focus on single channel speech enhancement framework, which is an important building block for such devices.

Over the past five decades, a vast amount of short-time spectral domain speech enhancement algorithms had been published and developed for applications such as mobile phones and hearing aids. In terms of single channel approach, the best known methods are the spectral subtraction (SS) (Boll, 1979; Berouti et al., 1979; Gustafsson et al., 2002), the minimum mean square error (MMSE) based estimator (Ephraim and Malah, 1984, 1985), and the Wiener filter (WF) (Scalart, 1996). Among these algorithms, SS is more often utilized in real world implementation due to its relative simplicity, which only requires an estimate of the noise power spectrum for computing the *a posteriori* signal-to-noise ratio (SNR). For MMSE based algorithms and WF method, *a priori* SNR is required, which involves an estimate of the clean speech signal. Although this increases the complexity of the problem, it was stated in the literature that the performance of the gain functions is mainly determined by the *a priori* SNR, while the *a posteriori* SNR acts only as a correction parameter

* Corresponding author. Address: 1/40 Marquis Street, Bentley, WA 6102, Australia. Tel.: +61 432160048.
E-mail address: peichee.yong@postgrad.curtin.edu.au (P.C. Yong).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.09.004

for low *a priori* SNR (Cappé, 1994). Since SS employs only the *a posteriori* SNR without utilizing the statistics and the distributions of the stochastic signal process, its performance is limited, which results in audible sound artifacts in the enhanced speech signal known as the musical noise. In order to solve this, a speech enhancement scheme in the modulation domain rather than in the conventional acoustic domain has been proposed in the literature (Paliwal et al., 2010, 2012). However, a practical solution to reduce the musical noise is by improving the *a priori* SNR estimate in the acoustic domain.

The most widely used approach for estimating the *a priori* SNR is the decision-directed (DD) approach (Ephraim and Malah, 1984). The DD approach performs a linear combination of two components: one being an estimate of previous *a priori* SNR and another being the maximum-likelihood (ML) SNR estimate. By applying a weighting factor close to unity of the past *a priori* SNR estimate, the DD approach corresponds to a highly smoothed version of the *a posteriori* SNR, which reduces the musical noise (Cappé, 1994). The drawback of reducing the variance in the *a priori* SNR estimate is that it cannot react quickly to abrupt changes in the instantaneous SNR. This leads to a performance degradation in speech enhancement scheme due to the speech transient distortion. In order to reduce the transient distortion, many algorithms have been proposed in the literature (Cohen, 2004; Chang et al., 2006; Plapous et al., 2006; Park and Chang, 2007; Breithaupt et al., 2008; Suhadi et al., 2011). Most of them have outperformed the traditional DD approach in terms of objective evaluations (Alam et al., 2009).

In general, the performance of a speech enhancement scheme depends on the joint temporal dynamics between the SNR estimate and the gain function (Breithaupt and Martin, 2011). For instance, the MMSE log-spectral amplitude (LSA) estimator with the DD *a priori* SNR estimate can generate speech signals without audible musical noise, provided that the weighting factor is close to unity (Ephraim and Malah, 1985; Loizou, 2007). Unlike the LSA approach, WF with the DD approach generates more speech distortion and musical noise. The main reason behind this is that the WF is a more aggressive gain function for low SNR. The result is a tendency to suppress more weak-speech components together with the residual noise. Thus, when compared to WF method, the LSA approach is preferred for less musical noise and speech distortion. In addition to that, much progress has been made in the development of MMSE estimators based on different cost functions and/or different statistical prior models to improve speech quality (Breithaupt and Martin, 2011; Breithaupt et al., 2008; Plourde and Champagne, 2009; Andrianakis and White, 2009). However, these algorithms involve the calculation of the confluent hypergeometric functions, which require a lot more computational complexity to implement when compared to WF and LSA methods.

Here, we are interested in developing a low complexity gain function, which employs the *a priori* SNR estimate with good noise suppression performance for real-time implementation. As such, the WF and LSA approaches will be used as the benchmark. Another method to obtain the gain function is to use a sigmoid (SIG) function. The rationale for using SIG function as a gain function is that it is a logistic function and can be viewed as a general CDF function (Yong et al., 2011). This gain function provides several parameters that can be adjusted to flexibly model exponential distributions. By optimizing the parameters of a SIG function, a well-balanced trade-off between noise reduction, speech distortion and musical noise can be achieved (Yong et al., 2011). Although this can also be achieved by employing an over-subtraction parameter on WF, only the mean of WF will be shifted while the shape will remain unchanged. This will give different sensitivity in the feedback when the gain is applied to the *a priori* SNR estimate. Therefore, a modified WF is not preferable as it does not provide as much flexibility offered by SIG function.

In this work, a modified sigmoid (MSIG) gain function has been proposed to increase the flexibility of the speech enhancement gain function and to provide a vehicle to enhance the speech quality in high noise conditions. The MSIG function combines a logistic function with a hyperbolic tangent function, providing a more flexible gain function with three controllable parameters. Optimization has been performed on these parameters to fit the MSIG function to either the LSA approach, WF method or SIG function to demonstrate the flexibility of the proposed gain function. In addition, a modified DD (MDD) SNR estimator has been developed, which basically nullifies the one frame delay in the conventional DD SNR estimator without increasing the computational complexity. This is achieved by matching the estimate of the clean speech spectrum and the *a priori* SNR estimate with the noisy speech spectrum in the current frame rather than the previous one. As a result, the MDD approach reduces the one-frame delay problem, which results in an improvement of speech quality. Extensive performance evaluations have been done between three MSIG functions, LSA, WF and SIG approaches, together with DD and MDD SNR estimates. Two objective measures, namely the perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and the segmental SNR (SNRseg) measures (Hansen and Pellom, 1998), together with a trade-off evaluation metric and the subjective listening tests were employed. Simulation results show the trade-off between the musical noise, the noise reduction and the speech distortion and demonstrate the advantages with the proposed MSIG functions and MDD SNR estimation.

The remainder of this paper is organized as follows. Section 2 gives a system overview. Section 3 shows the proposed SNR estimate. Section 4 demonstrates the modified gain function. Section 5 outlines the objective measures used for performance evaluation and Section 6 presents the results. Section 7 concludes the paper.

2. System overview

The goal of speech enhancement is to compute the enhanced speech signal $\hat{x}(n)$, given a noisy signal $y(n) = x(n) + v(n)$, where $x(n)$ is the clean speech signal and $v(n)$ is the uncorrelated additive noise. By using the short-time Fourier transform (STFT), the spectral components of the noisy speech $Y(k,m)$ can be obtained by

$$Y(k,m) = \sum_{n=1}^{N} y(mR+n)w(n)\exp\left(\frac{-j2\pi kn}{N}\right), \tag{1}$$

where k is the frequency bin index, m is the frame index, R is the frame rate and $w(n)$ is a window function. The clean speech spectrum estimate $\widehat{X}(k,m)$ is then obtained by

$$\widehat{X}(k,m) = G(k,m)Y(k,m) \tag{2}$$

where $G(k,m)$ is a non-linear spectral gain function. The gain function can be expressed as a function of the *a priori* SNR

$$\xi(k,m) = \frac{E\left\{|X(k,m)|^2\right\}}{E\left\{|V(k,m)|^2\right\}} = \frac{\lambda_x(k,m)}{\lambda_v(k,m)} \tag{3}$$

and/or the *a posteriori* SNR

$$\gamma(k,m) = \frac{|Y(k,m)|^2}{E\left\{|V(k,m)|^2\right\}} = \frac{|Y(k,m)|^2}{\lambda_v(k,m)}, \tag{4}$$

where $\lambda_x(k,m)$ and $\lambda_v(k,m)$ denote clean speech power spectral density (PSD) and noise PSD, respectively.

The gain function is often derived from MMSE optimization criteria. One of those is the WF method, which minimizes the expected value $\mathrm{E}\{|X(k,m) - \widehat{X}(k,m)|^2\}$. It can be computed using the *a priori* SNR as

$$G_{\mathrm{WF}}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}. \tag{5}$$

Other widely used algorithms are based on a direct estimate of the clean speech spectral magnitude. One of them is the LSA estimator, which can be obtained by minimizing $\mathrm{E}\{[\log(|X(k,m)|) - \log(|\widehat{X}(k,m)|)]^2\}$ (Ephraim and Malah, 1985). The resulting gain function for the LSA approach can be obtained as

$$G_{\mathrm{LSA}}(k,m) = \min\left\{\varsigma, \frac{\xi(k,m)}{1+\xi(k,m)}\left[\frac{1}{2}\int_{v(k,m)}^{\infty}\frac{e^{-t}}{t}dt\right]\right\} \tag{6}$$

with

$$v(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)}\gamma(k,m) \tag{7}$$

where ς denotes a practical upper bound used to prevent a large gain value at low *a posteriori* SNR. Here, we choose $\varsigma = 10$.

3. A priori SNR estimation

Prior to the computation of the spectral gain function, two parameters have to be estimated: the noise PSD and the *a priori* SNR. In this work, the noise PSD is estimated by using the MMSE noise estimator in (Hendriks et al., 2010). Meanwhile for the estimation of the *a priori* SNR, the most widespread method is the DD approach, given by Ephraim and Malah (1984)

$$\hat{\xi}_{\mathrm{DD}}(k,m) = \max\left\{\beta\frac{|\widehat{X}(k,m-1)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1], \xi_o\right\}, \tag{8}$$

where $\hat{\lambda}_v(k,m)$ and $\widehat{X}(k,m-1)$ denote, respectively, the estimated noise PSD and the estimated clean speech spectrum from the preceding frame. The parameter β denotes the smoothing factor, $P[.]$ denotes the half-wave rectification and ξ_o denotes a SNR floor. As observed from the equation, the first term is the *a priori* SNR estimate in the previous frame while the second term is an ML estimate computed from the *a posteriori* SNR. The advantage of the DD approach is its capability to eliminate musical noise based on the choice of β in the conditional smoothing procedure (Cappé, 1994). It was suggested to set β close to unity so that the musical noise can be eliminated, particularly in conjunction with the MMSE estimator approach. However, this leads to a slow update of the *a priori* SNR estimate, resulting in speech transient distortion, especially in speech onsets. This is due to little influence of the second term $(1-\beta)P[\gamma(k,m)-1]$ in the update.

In addition, the DD approach based on Eq. (8) has an extra frame delay during speech transients since it employs the previous frame clean speech spectrum estimate. As a consequence, the gain function matches the previous frame instead of the current one. Although the delay can be reduced by choosing a smaller β in Eq. (8), more musical noise will be perceived since $(1-\beta)P[\gamma(k,m)-1]$ is usually unsmoothed. Thus, we propose a MDD approach to reduce the delay in speech transients by matching both estimates of the clean speech spectrum and the *a priori* SNR estimate with the current noisy speech spectrum. The first term of the DD approach is modified such that the gain function at previous frame is mapped with the current noisy speech spectrum rather than the previous one. As such, the MDD approach is given by

$$\hat{\xi}_{\mathrm{MDD}}(k,m) = \max\left\{\beta\frac{|G_{(.)}(k,m-1)Y(k,m)|^2}{\hat{\lambda}_v(k,m)} + (1-\beta)P[\gamma(k,m)-1], \xi_o\right\} \tag{9}$$

where $G_{(.)}$ indicates that the same gain function is used to obtain both the *a priori* SNR estimate and the speech estimate. The advantage of this new approach is that it has the same computational complexity as the DD approach while having a better enhanced speech quality, which makes it suitable for real-time implementation.

According to Eq. (9), the first term of the MDD approach does not contain an estimate of the *a priori* SNR at previous frame when compared to the original method. Therefore, the MDD approach is no longer representing a conditional first order recursive averaging algorithm as in Eq. (8). This means that it increases the sensitivity of the *a priori* SNR estimate towards abrupt changes in speech signal, which directly reduces the speech transient distortion. However, such variance in the *a priori* SNR estimate can lead to audible musical noise due to the higher sensitivity to changes. In order to reduce, or eliminate such musical noise, a first order recursive smoothing procedure can be applied in the *a posteriori* SNR estimation in Eq. (4) as Yong et al. (2011)

$$\bar{\gamma}(k,m) = \frac{\lambda_y(k,m)}{\hat{\lambda}_v(k,m)} \tag{10}$$

where

$$\lambda_y(k,m) = \alpha_y \lambda_y(k,m-1) + (1-\alpha_y)|Y(k,m)|^2. \tag{11}$$

The parameter λ_y is the noisy speech PSD, which is obtained by smoothing the magnitude square of the noisy signal. The smoothing constant is defined as $\alpha_y = \exp\left(\frac{-2.2R}{t_y f_s}\right)$, where R is the frame rate from Eq. (1), while t_y and f_s denote the time averaging constant and the sampling rate, respectively.

4. Modified sigmoid gain function

Most of the gain functions developed for speech enhancement scheme are based on optimization criteria, such as the LSA approach, the WF method and all other MMSE-based algorithms (Breithaupt et al., 2008; Plourde and Champagne, 2009; Andrianakis and White, 2009). The problem is that the optimization of the criteria is made under certain model conditions such as stationarity and certain distributions. Ideally it is desirable to have a gain function that offers optimal performance in all scenarios. This will lead to different cost functions and gain functions giving different performance in different background noise scenarios. Apart from that, some of them involve complex mathematics equations that require large computational load to solve, making them sometimes less attractive in real world scenarios. Thus, we propose to design a gain function with low complexity and high flexibility, while having similar or better performance when compared to most of the MMSE estimators. The important consideration is to have control over the shape of the gain function. For this purpose, a flexible sigmoid-shape function is utilized. By designing the SIG function with different shapes, a similar performance as the MMSE-based estimators will be obtained. The rationale behind using the SIG function is that it is a general CDF function with a shape that can be adjusted by several tunable parameters. In that case, the quality of the enhanced speech can be improved.

In previous work, a SIG function has been presented to map with the *a posteriori* SNR in (Yong et al., 2011). However, instead of mapping with the *a posteriori* SNR, which limits the performance of the gain function, here the SIG function is mapped with the *a priori* SNR estimate. The gain function is given as

$$G_{\text{SIG}}(k,m) = \frac{1}{1+\exp\left[-a\left(\hat{\xi}(k,m)-c\right)\right]} \tag{12}$$

where a and c are parameters that control the slope and the mean of the gain curve, respectively. Both parameters control the amount of musical noise, speech distortion and noise reduction in the enhanced speech. In order to obtain a balanced trade-off between them, the sigmoid slope has to be sensitive towards speech and less sensitive towards the variation of noise. In this case, the mean of the SIG function has to be less than 1. This is not plausible as when the mean value is approaching zero, the gain value will not reach zero until a very small SNR value, which leads to insufficient noise reduction. To provide more noise reduction at low SNR conditions, a MSIG function is developed, which has three parameters that can be adjusted or optimized for better enhanced speech quality. The proposed function is obtained by multiplying the original logistic function Eq. (12) with a hyperbolic tangent function, as

$$G_{\text{MSIG}}(k,m) = \frac{1-\exp\left[-a_1\hat{\xi}(k,m)\right]}{1+\exp\left[-a_1\hat{\xi}(k,m)\right]} \times \frac{1}{1+\exp\left(-a_2\left[\hat{\xi}(k,m)-c\right]\right)}. \tag{13}$$

By changing the parameter values, the behaviour of the MSIG function can be made similar to the different conventional gain functions, such as the LSA, the WF and the SIG approaches. To achieve this an optimization problem has been set up. The problem can be formulated as

$$\min_{\mathbf{z}} \|G_{\text{MSIG}}(\mathbf{z},x) - D(x)\|_2^2 \tag{14}$$

where $\mathbf{z} = [a_1\ a_2\ c]$ and $D(x)$ is a gain function chosen from WF, LSA and SIG. The optimization problem in Eq. (14) is a non-linear optimization problem in terms of the parameter $\mathbf{z}$. A solution for the problem can be obtained by using the minimization function *lsqnonlin* in MATLAB, which solves the non-linear least-square curve fitting problem by using a trust region reflective Newton method. As such, MSIG parameters that best fit the gain function in $D(x)$ in the least-square sense can be found in Section 6.

5. Representative objective measures

Many objective measurement algorithms have been derived in the literature for evaluating the performance of speech enhancement algorithms (Loizou, 2007; Hu and Loizou, 2008). The most widely used methods include the PESQ measure (Rix et al., 2001) and the SNRseg measure (Hansen and Pellom, 1998). The PESQ measure, which was not originally designed to evaluate the performance of speech enhancement algorithms, has been found to have a good correlation overall with mean opinion score (MOS) (Hu and Loizou, 2008). It predicts the MOS scores which yields a result from 1 to 5, where a higher score indicates a better speech quality. Meanwhile, the SNRseg measure is also preferred among the vast amount of objective measures as it has been found to correlate best with background noise reduction (Hu and Loizou, 2008).

In this paper, both the PESQ measure and the SNRseg measure were used to evaluate the performance of the proposed algorithms. The PESQ measure was implemented based on the procedures presented in (Loizou, 2007). The SNRseg measure is defined as Loizou (2007)

$$\mathrm{SNRseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{||\mathbf{x}(m)||^2}{||\mathbf{x}(m)-\hat{\mathbf{x}}(m)||^2} \tag{15}$$

where the vector $\mathbf{x}(m)$ represents a clean speech (time-domain) frame, and $\hat{\mathbf{x}}(m)$ is the enhanced speech frame. In order to discard non-speech frames, each frame was threshold by a -10 dB lower bound and a 35 dB upper bound.

The performance of the speech enhancement scheme has a trade-off between musical noise, speech distortion and noise reduction. The PESQ measure and the SNRseg measure can not represent the whole picture of these trade-offs. Therefore, an objective evaluation metric is also utilized to evaluate and compare the results between the amount of musical noise, speech distortion and noise reduction generated from the speech enhancement scheme.

First of all, the musical noise and the noise reduction should only be calculated during noise-only periods in short-time spectral domain. Since in practical situations the true noise PSD is often not known, a reference VAD for the clean speech is required for performance evaluation without the knowledge of noise characteristics. In order to obtain the VAD decisions at different frames and frequency bins, a multi decisions sub-band VAD (MDSVAD) is utilized (Davis et al., 2006). Given two hypotheses, $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$, which indicate speech absence and presence respectively in the kth frequency bin of the mth frame, the MDSVAD is given by

$$D(k,m) = \begin{cases} 1 & \text{if } \mathcal{H}_0(k,m) \\ 0 & \text{if } \mathcal{H}_1(k,m). \end{cases} \tag{16}$$

The amount of musical noise is believed to be highly correlated with the number of isolated spectral components and their level of isolation (Uemura et al., 2008). Since such components have relatively high power, they can be perceived as tonal sound that is strongly related to the weight of skirt of the probability density function (PDF). A signal with skirt can be identified using higher-order statistics, such as kurtosis. However, in order to identify only the musical-noise components, a kurtosis ratio (KurtR) is used to measure the change in kurtosis between the noisy signal and the enhanced signal. In (Uemura et al., 2008), this ratio was derived as a function controlled by the over-subtraction factor in the SS function as well as the shape parameters from the distribution model of the speech or noise signal. The KurtR in this paper is determined by the actual noisy speech signal and the enhanced speech signal during noise-only periods. Such measure is defined as

$$\mathrm{KurtR} = E\left\{\frac{\mathcal{K}_{\hat{x}}(k)}{\mathcal{K}_y(k)}\right\} \tag{17}$$

where $\mathcal{K}_{\hat{x}}(k)$ and $\mathcal{K}_y(k)$ denote the kurtosis of the enhanced signal and the noisy signal, respectively at k-th frequency bin. Both of them are computed only during speech absence periods, as given by

$$\mathcal{K}_{\hat{x}}(k) = \frac{\sum_{m=0}^{M-1}|\widehat{X}(k,m)D(k,m)|^4}{\left[\sum_{m=0}^{M-1}|\widehat{X}(k,m)D(k,m)|^2\right]^2} - 2. \tag{18}$$

$$\mathcal{K}_y(k) = \frac{\sum_{m=0}^{M-1}|Y(k,m)D(k,m)|^4}{\left[\sum_{m=0}^{M-1}|Y(k,m)D(k,m)|^2\right]^2} - 2. \tag{19}$$

A smaller value of KurtR in Eq. (17) indicates less musical noise.

Meanwhile, the amount of noise reduction can be defined as the input noise power in dB minus the output noise power in dB. This noise reduction ratio (NRR) is defined during noise-only periods as

$$\mathrm{NRR}[dB] = 10\log_{10}\frac{\sum_{m=0}^{M-1}\sum_{k=1}^{K}|Y(k,m)D(k,m)|^2}{\sum_{m=0}^{M-1}\sum_{k=1}^{K}|\widehat{X}(k,m)D(k,m)|^2}. \tag{20}$$

For speech distortion measure, the log-likelihood (LLR) measure is used. It is a spectral distance measure that models the mismatch between the formants of the clean and enhanced speech signals (Quackenbush et al., 1988). The LLR measure is defined as Quackenbush et al. (1988)

$$d_{\mathrm{LLR}}\left(\vec{l}_{\hat{x}}, \vec{l}_x\right) = \frac{\vec{l}_{\hat{x}}\mathrm{R}_x\vec{l}_{\hat{x}}^T}{\vec{l}_x\mathrm{R}_x\vec{l}_x^T} \tag{21}$$

where $\vec{l}_x$ and $\vec{l}_{\hat{x}}$ are the linear predictive coding (LPC) coefficient vectors of the clean speech signal and the enhanced speech signal respectively, and R_x is the autocorrelation matrix of the clean speech signal. A lower LLR score indicates a better speech quality.

The objective evaluation metric provides a multi-criteria evaluation for the various parameters that define the speech enhancement methods. This helps to identify reasonable parameters and to provide an indication of parameter sensitivity. As such the objective evaluation has been used to obtain reasonable candidates for listening tests.

6. Experimental evaluation

6.1. Parameter optimization of the MSIG function

The parameters of the MSIG function were fitted to either WF, LSA or SIG function by using Eq. (14). For the SIG function, the parameters chosen were $a_2 = 1$ and

Table 1
MSIG parameters.

Functions	Parameters			Fitted curve
	a_1	a_2	c	
MSIG-fix1	2.3918	0.2120	−1.7071	LSA
MSIG-fix2	11.6869	0.4337	0.7556	WF
MSIG-fix3	15	0.6351	0.2243	SIG

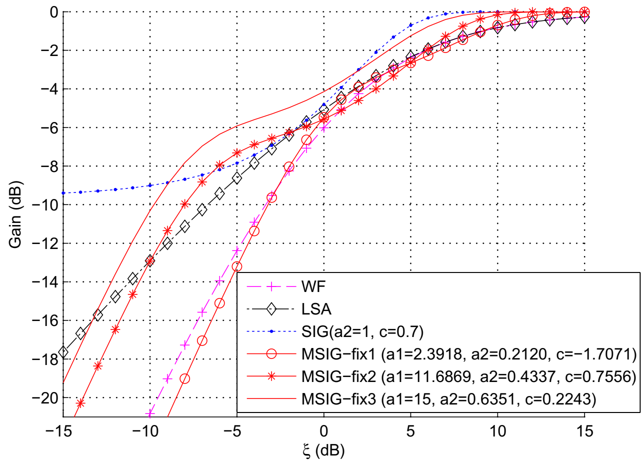

Fig. 1. Gain curves of different algorithms, as functions of the *a priori* SNR $\xi(k,m)$, where $\gamma(k,m) = \xi(k,m) + 1$.

$c = 0.7$. In the optimization procedure, the initial estimates of MSIG parameters were set as $\mathbf{z}_0 = [0\ 0\ 0]$. An upper bound constraint was employed in the optimization procedure, such that $\mathbf{z}_{ub} = [15\ 1\ 1]$. The curve fitting was done under the condition that the instantaneous SNR is equal to the *a priori* SNR. The results of the curve fitting can be observed in Table 1, which shows the optimized parameters for three different MSIG curves from Eq. (13).

Fig. 1 plots the MSIG curves as functions of the *a priori* SNR, together with the WF gain in Eq. (5), the LSA approach in Eq. (6) and the SIG function from Eq. (12). From the figure, MSIG-fix1 is fitted to WF, but with slightly higher attenuation at low SNR conditions (below −4 dB). Also, MSIG-fix2 is fitted to the LSA estimator but is a more aggressive gain function below −10 dB SNR. The least aggressive gain function can be found in MSIG-fix3, which does not really match the SIG function. This is because of the upper bound constraint of the parameter a_1, which offers sufficiently small gain values at lower SNR region. If the upper bounds were not imposed, MSIG will fully fit the SIG function at $a_1 = 92043$, $a_2 = 1$ and $c = 0.7$. An advantage of all the MSIG functions over the conventional methods is a lower gain value at low SNR, while having a larger gain value at high SNR region. This allows more noise to be suppressed and more speech components to be preserved.

6.2. Experimental setup

For objective evaluation, 30 IEEE speech sequences were taken from NOIZEUS speech database (Loizou, 2007) and were added with pink noise. The tests were done with 0.01 step for both $0 \leqslant t_x \leqslant 0.1$ and $0.9 \leqslant \beta \leqslant 0.99$. The smoothing constant α_y was plotted instead of t_x, in conjunction to β for consistency in terms of the frame rate. The reference MDSVAD in Eq. (16) were generated from the same speech sequences but with 50 dB global SNR to reduce miss-detections of speech. All results were generated with $K = 512$ frequency bins. A square-root Hann window was used for $w(n)$ with 50% overlap ($R = 256$). The value of the *a priori* SNR floor was chosen as $\xi_o = -25$ dB. In addition, a constant residual noise floor, $\epsilon = -15$ dB was employed for all the gain functions, such that

$$G_{(\cdot)}(k,m) = \max\left\{\epsilon, G_{(\cdot)}(k,m)\right\}. \tag{22}$$

The results of the performance evaluation will differ with the noise estimate $\hat{\lambda}_v(k,m)$. In this experiment, the results were generated with the MMSE noise estimator in (Hendriks et al., 2010). As such, a consistent simulation can be run, where the only changes in the system are the gain functions and the SNR estimators.

The performance of the proposed approach is further verified by subjective listening tests, where the listeners provide ratings for each individual component of a noisy speech signal, which include the speech signal, the background noise, and the musical noise (Hu and Loizou, 2008). The listener was prompted to rate the noisy and the enhanced speech signal on the following three criteria:

1. SPCH: the speech signal using a 5 point scale of signal distortion;
2. NSE: the noise using a 5 point scale of background intrusiveness;
3. MUSIC: the musical noise using a 5 point scale of musicalness.

Table 2
Description of the SPCH, NSE and MUSIC scales used in the listening tests.

Rating	Description
SPCH	
5	No degradation, very natural
4	Little degradation, fairly natural
3	Somewhat degraded, somewhat natural
2	Fairly degraded, fairly unnatural
1	Very degraded, very unnatural
NSE	
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive
MUSIC	
5	Not perceptible
4	Somewhat perceptible
3	Perceptible but not annoying
2	Fairly conspicuous, somewhat annoying
1	Very conspicuous, very annoying

The SPCH, NSE and MUSIC scales are described in Table 2. Note that those numbers are not absolute scales but serve as an indication of hearing experience to evaluate speech quality. A total of eight listeners (six males, two females aged between 20 and 30) were recruited for the listening tests. Five separate sentences, which consist of three male spoken sentences and two female spoken sentences, were included for the tests.They were taken from the same NOIZEUS database used for objective evaluation. Each of them were corrupted with either pink noise or factory noise, at 0 dB and 15 dB SNRs. Tests were performed by using audio-technica ATH-ESW9 headphones. The subjects were categorized into two groups: one group was required to start the tests with 0 dB SNR cases, while the other group began the tests from 15 dB SNR cases. During the tests, the listeners were not given information about the type of the gain function and the method for the SNR estimate used in each utterance. The listeners were allowed to listen to each utterance several times with access to the clean speech signal and noisy signal references. The average duration of a test was approximately 2 hours per subject.

6.3. Evaluation of the proposed MDD in estimating a priori SNR

The performance of the MDD approach in estimating the *a priori* SNR is compared with the DD approach and a reference method from Park and Chang (2007). Instead of using a fixed smoothing factor β, the reference method, ξ_{ref} modifies the DD approach by using a sigmoid function to control the weighting value for β. Fig. 2 shows an example of the *a priori* SNR estimation for $\xi_{\text{DD}}, \xi_{\text{ref}}$ and ξ_{MDD} approaches with different gain functions when speech is detected. It can be observed that the *a priori* SNR for all the gain functions can be represented as the smoothed version of the *a posteriori* SNR. For the conventional DD approach, $\hat{\xi}_{\text{DD}}(k,m)$ follows the $\gamma(k,m)-1$ estimate with one frame delay in speech frames when β is close to 1, ($\beta=0.98$). Both the reference method and the MDD approach improve the DD approach in terms of reducing the delays in speech onsets. However, when the speech stops, ξ_{MDD} follows the *a posteriori* SNR estimate but ξ_{ref} gets a one frame delay. Thus, the proposed method is superior in estimating the *a priori* SNR. It is a direct yet effective solution to reduce and eliminate the distortion at speech transients. The same patterns of improvement can be seen for all the evaluated gain functions. In addition, the proposed method has potentially lower computational complexity when compared to the reference method in (Park and Chang, 2007), which is beneficial for many real-time applications.

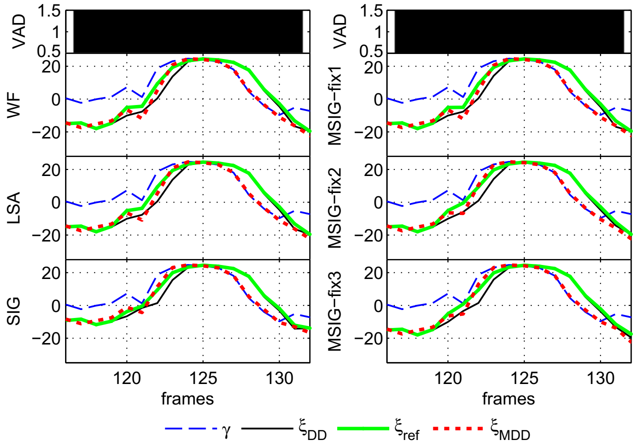

Fig. 2. Comparison between MDSVAD decisions, $\gamma(k,m)-1$ (blue dashed line), $\hat{\xi}_{\text{DD}}(k,m)$ (black solid line), $\hat{\xi}_{\text{ref}}(k,m)$ (green solid line) and $\hat{\xi}_{\text{MDD}}(k,m)$ (red dotted line) at 937.5 Hz and 15 dB SNR. Here, $\beta=0.98$ were applied for $\hat{\xi}_{\text{DD}}(k,m)$ and $\hat{\xi}_{\text{MDD}}(k,m)$, while $\alpha_y=0.3$ were employed for all evaluated *a priori* SNR estimators. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.4. Objective performance evaluation

Evaluation is performed for both DD and MDD SNR estimators, for different gain functions, which include the WF, LSA, SIG, MSIG-fix1, MSIG-fix2, and MSIG-fix3 methods. The measurement employed are the PESQ measures, the SNRseg measures, and the evaluation metric which include the KurtR, NRR and LLR measures. For PESQ, SNRseg and NRR, higher scores indicate better results and better speech quality. Meanwhile, lower KurtR and LLR scores mean less musical noise and less speech distortion, respectively.

6.4.1. PESQ evaluation

Figs. 3–6 show the PESQ results for the DD and MDD approaches, respectively, for 0 dB and 15 dB SNRs. The PESQ scores obtained from the WF, LSA and MSIG-fix1 gain functions have a similar trend. They have better results at small values of β and α_y, while having performance drop when both β and α_y are increasing. This is because speech starts to sound unnatural and degraded when more smoothing is applied to the WF, LSA and MSIG-fix1 approaches. In addition, the decreasing rate of the PESQ scores for the MDD approach for an increasing β and α_y is slower than the DD approach, which results in a better PESQ results for the MDD approach when both β and α_y are large values.

For the SIG function, while the DD approach follows the previously described trend, the MDD approach has different trend. At 0 dB SNR, the SIG function with the MDD approach has better PESQ results for all β values at a large α_y. While at 15 dB SNR, it has the optimal PESQ scores at a smaller α_y, of which the values are the same for all β. This indicates that even with the same values of β and α_y, the amount of smoothing varies for different gain functions. Since the MDD approach provides better PESQ scores at most β values, particularly for $\beta > 0.94$, it is the preferred choice for SNR estimate when compared to the DD approach.

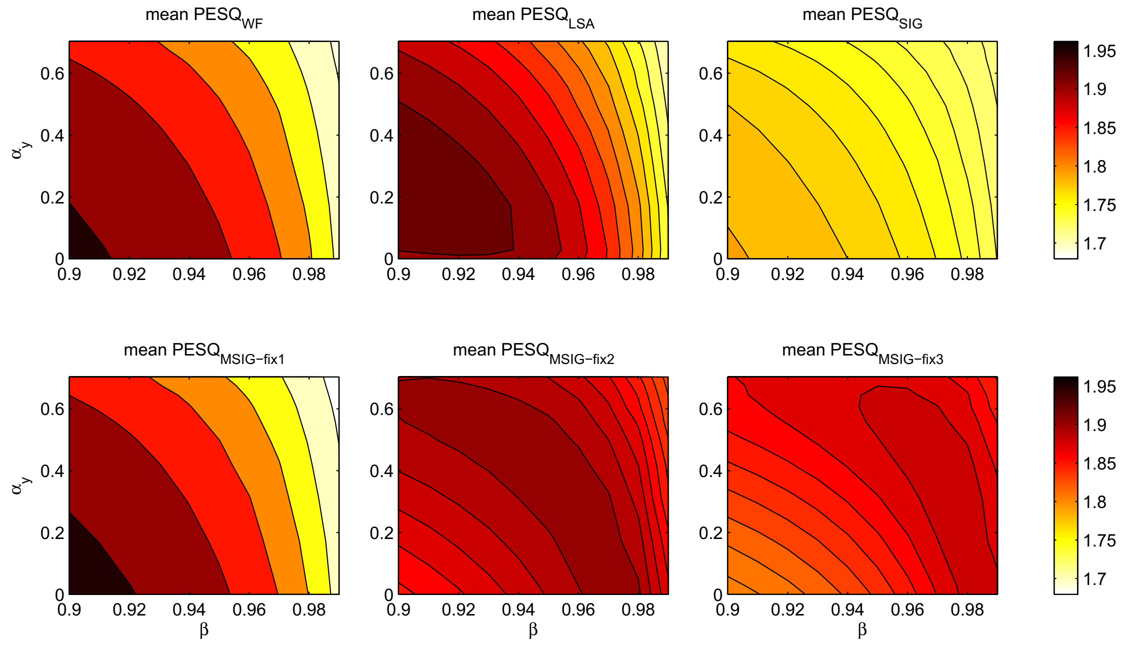

Fig. 3. Average PESQ scores with $\hat{\xi}_{DD}$ at SNR = 0 dB.

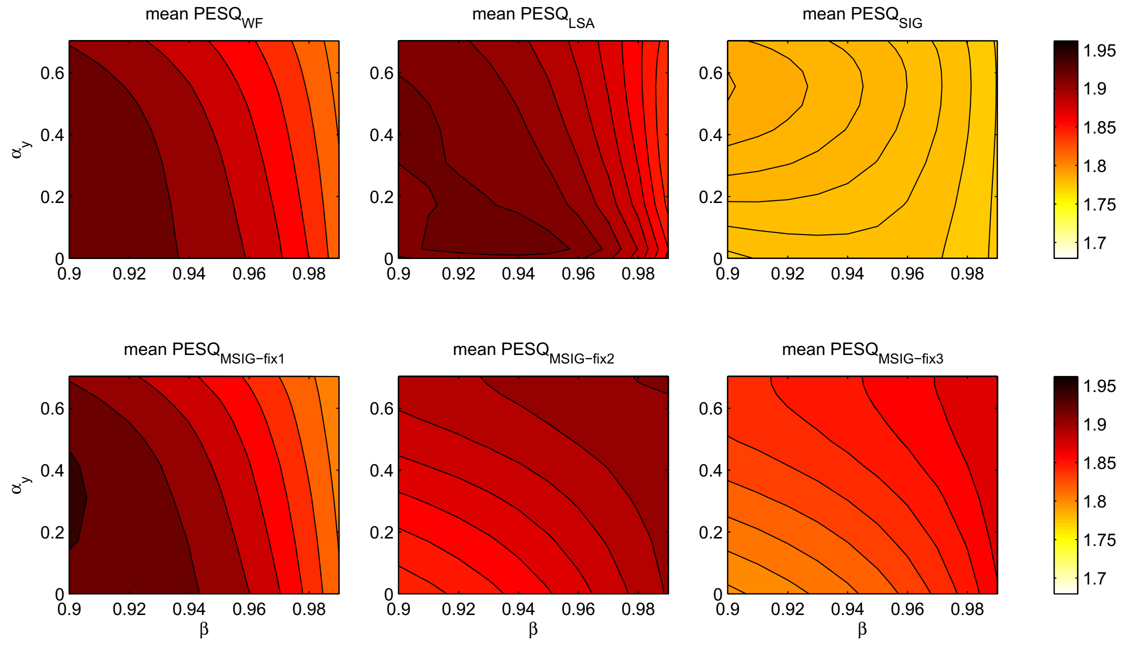

Fig. 4. Average PESQ scores with $\hat{\xi}_{MDD}$ at SNR = 0 dB.

The contour shape obtained from MSIG-fix2 is similar to MSIG-fix3, but totally different from WF, LSA, MSIG-fix1 and SIG. Both of them have better PESQ results recorded when β and α_y are large. This is because both MSIG-fix2 and MSIG-fix3, together with the SIG function, are non-aggressive gain functions. As shown in Fig. 1, they do not provide much noise suppression at low SNR. Therefore, by providing more smoothing to the SNR estimates for these two gain functions helps to reduce noise variations, which leads to better PESQ scores. In terms of the comparison between the SNR estimates, MSIG-fix2 has better PESQ scores in conjunction with the MDD approach when compared to the DD approach, while MSIG-fix3 has better PESQ scores when the DD approach is employed.

As observed from the figures, both the WF and MSIG-fix1 have the best PESQ results among all gain functions, with the MDD approach having the least parameter sensitivity over a wide range of parameters and the best scores for 15 dB global SNR conditions. By taking this into

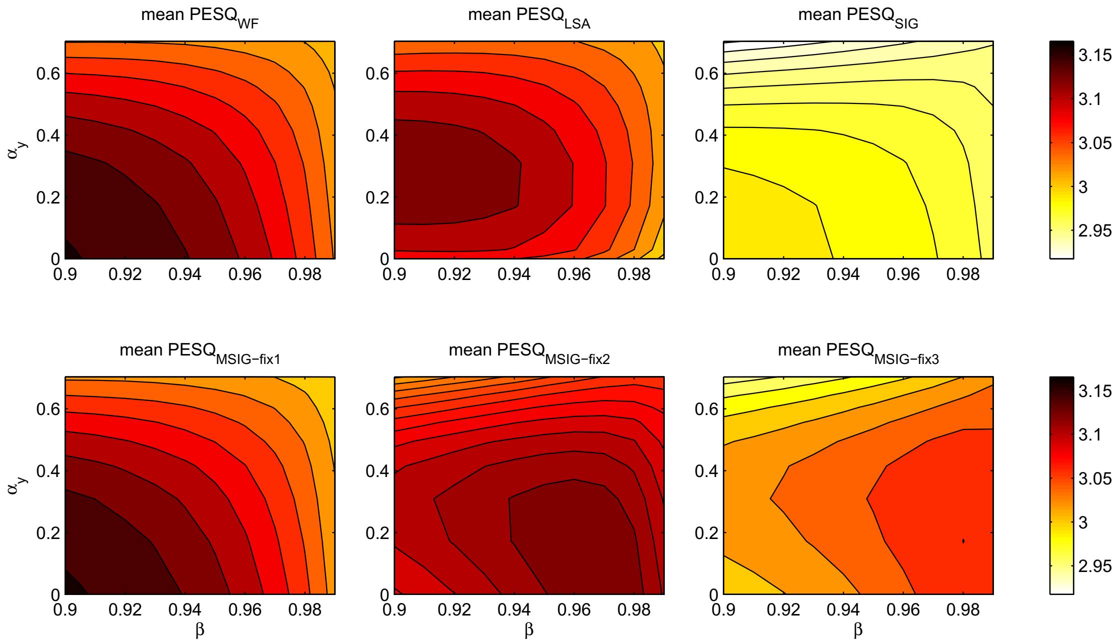

Fig. 5. Average PESQ scores with $\hat{\xi}_{DD}$ at SNR = 15 dB.

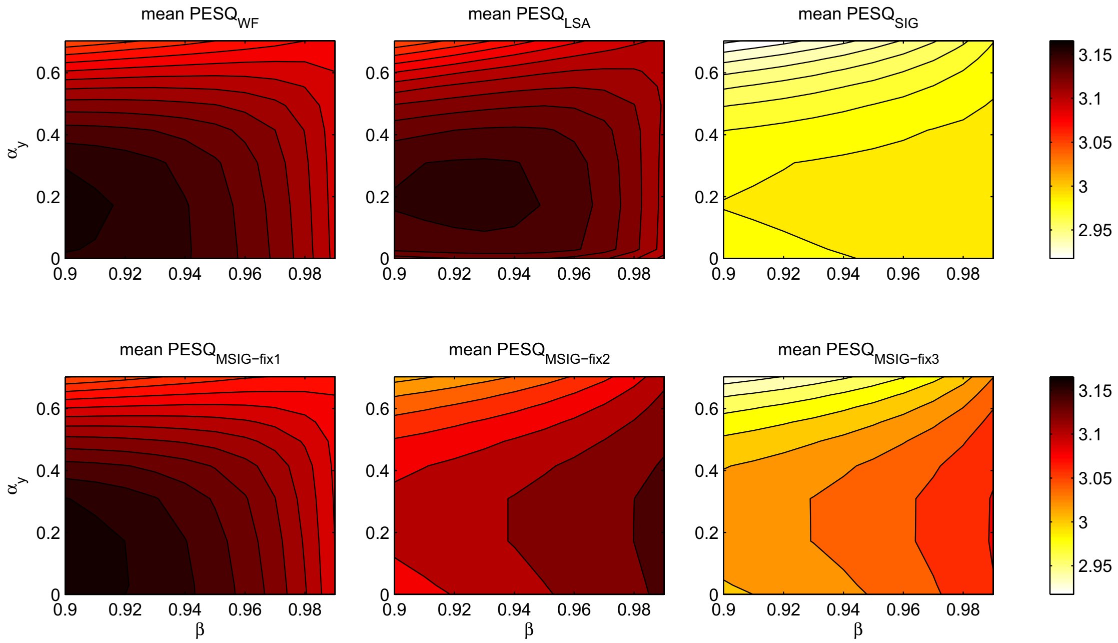

Fig. 6. Average PESQ scores with $\hat{\xi}_{MDD}$ at SNR = 15 dB.

account with the observation that better PESQ results have been obtained for the MDD approach for increasing β and α_y, MDD performs better than DD, particularly for WF, LSA and MSIG-fix1 in terms of the PESQ measure.

6.4.2. SNRseg evaluation

Figs. 7–10 present the SNRseg results for the DD and MDD approaches with 0 dB and 15 dB SNRs. All evaluated gain functions give better SNRseg results for the MDD approach when compared with the DD approach for large value of β. In particular, the MDD approach has a significant improvement over the DD approach for the WF, LSA, MSIG-fix1 and SIG gain functions for all β and α_y. This indicates that the segmental SNR increases when the delay in speech transients is reduced and removed. Also, the WF and MSIG-fix1 perform best when β and α_y are small, while LSA has better SNRseg results at high smoothing setting.

For MSIG-fix2 at 0 dB SNR, apart from having the optimal point at different smoothing parameters, the SNR-

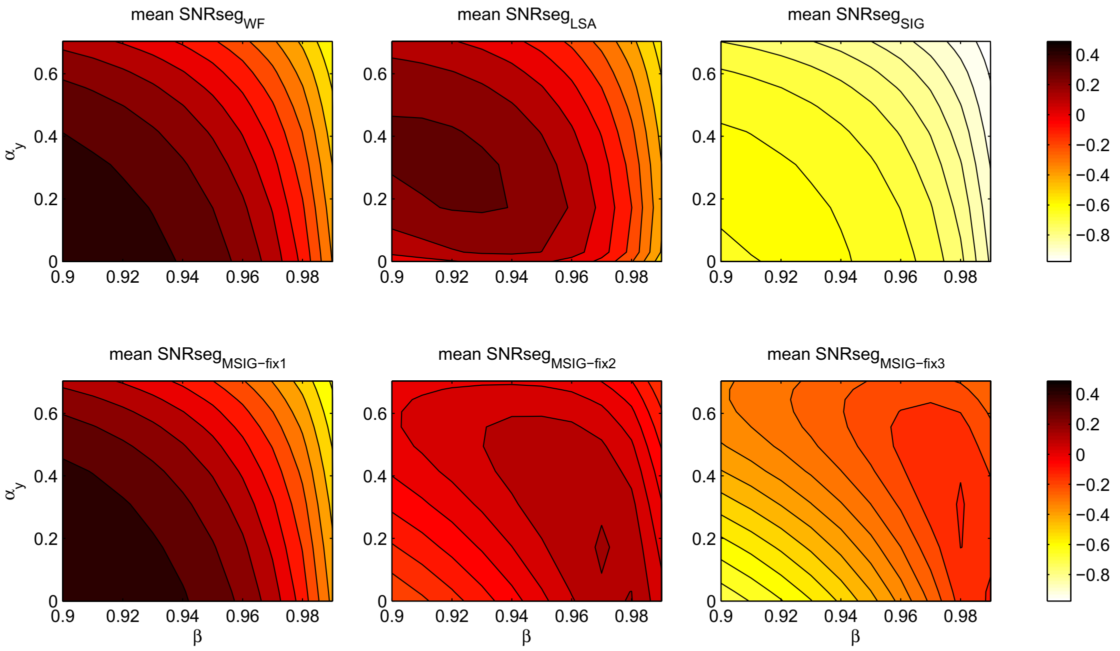

Fig. 7. Average SNRseg values with $\hat{\xi}_{\mathrm{DD}}$ at SNR = 0 dB.

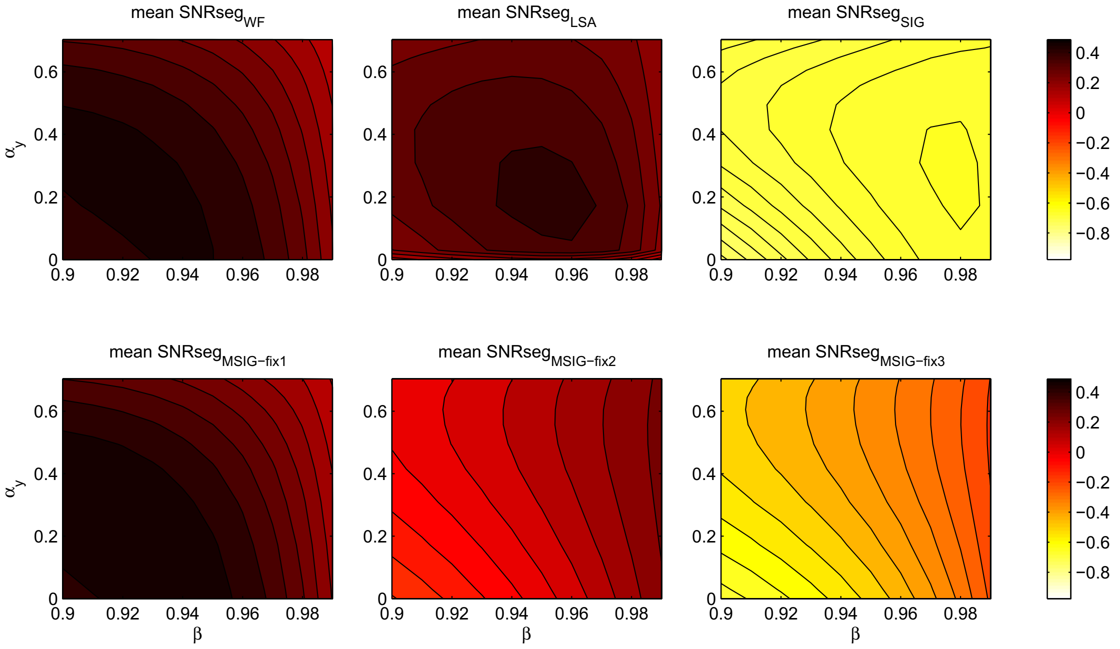

Fig. 8. Average SNRseg values with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 0 dB.

seg results for both DD and MDD approaches are very similar. For MSIG-fix3, it has poorer SNRseg scores for the MDD approach at 0 dB when compared to the DD approach. Despite that, at 15 dB SNR both MSIG-fix2 and MSIG-fix3 give better SNRseg results for the MDD approach. For SIG function, the MDD approach has similar performance with the DD approach when less smoothing is applied, but it becomes better than the DD approach for increasing β and α_y.

6.4.3. Objective speech distortion, musical noise and noise suppression evaluation

Figs. 11–14 present the results from the objective evaluation metric for the DD and MDD approaches at 0 dB and 15 dB SNRs, respectively. In terms of the amount of musical noise generated, KurtR decreases with increasing values of β and α_y for all evaluated gain functions with one exception (MSIG-fix3). The MSIG-fix3 function is a rather flat function over a range of input SNRs that lacks of

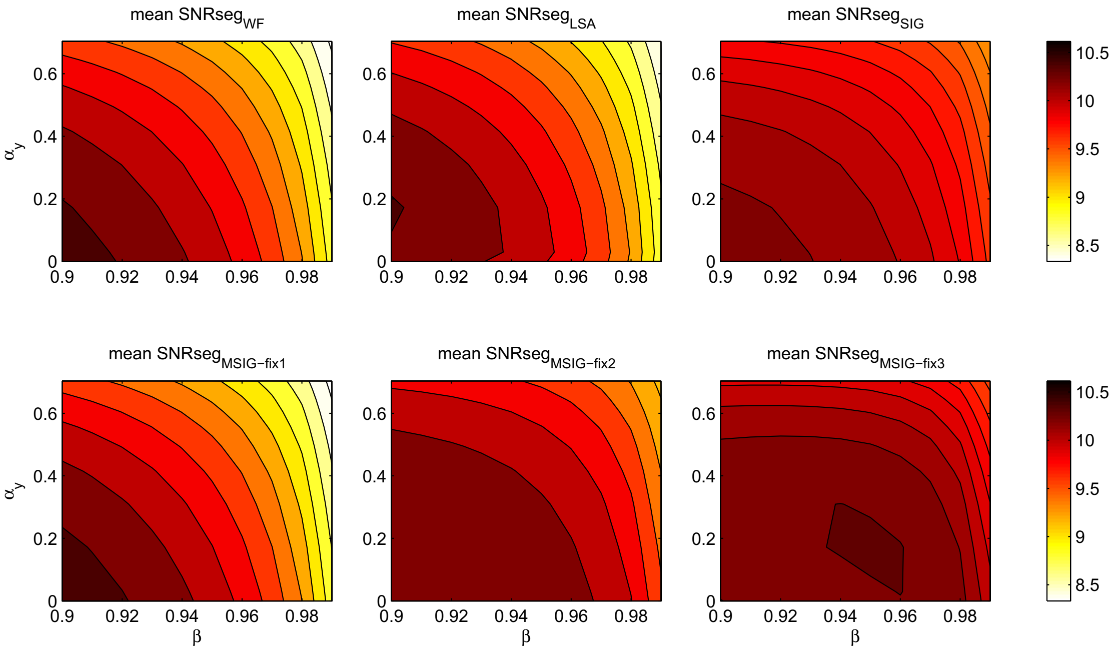

Fig. 9. Average SNRseg values with $\hat{\xi}_{DD}$ at SNR = 15 dB.

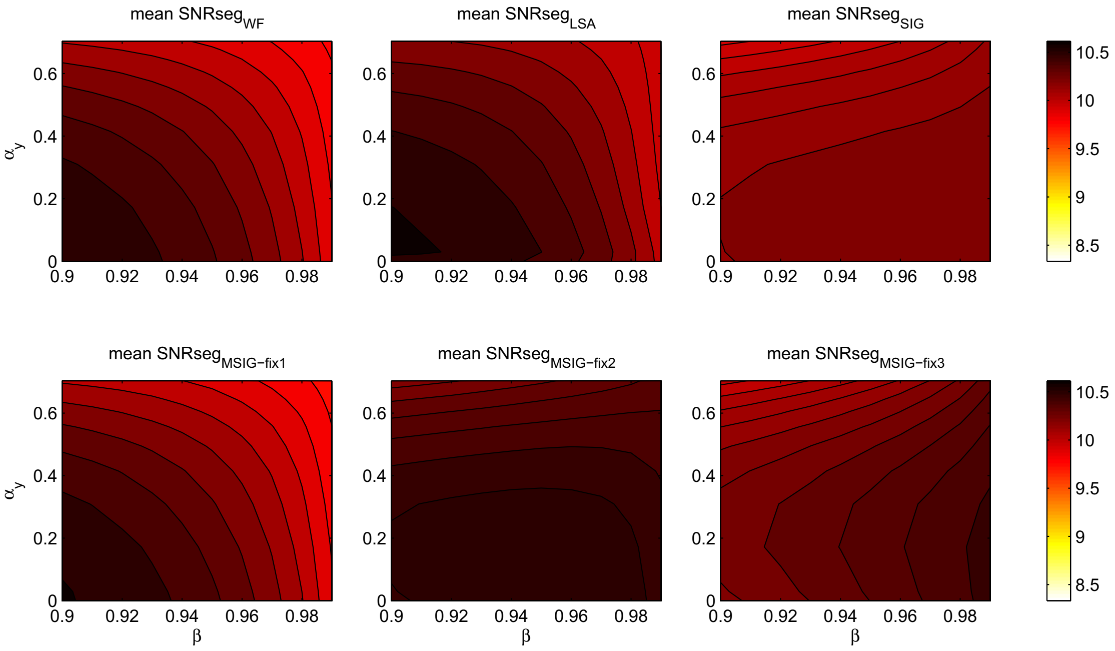

Fig. 10. Average SNRseg values with $\hat{\xi}_{MDD}$ at SNR = 15 dB.

distinctiveness which gives poor performance in terms of musical noise. For the rest of the evaluated gain functions, the DD approach performs better than MDD for KurtR with a few exceptions, which is due to the higher sensitivity to changes in MDD that provides capability to track speech onsets. For WF and MSIG-fix1, both the DD and MDD approaches have almost identical results for the KurtR measure at $\beta > 0.98$. Meanwhile, at 15 dB SNR, there is a significant improvement in performance for the MDD approach over the DD approach in conjunction with both WF and MSIG-fix1 functions when β is large. These results are the best among all evaluated gain functions. Since less musical noise is generated with large smoothing parameters, the MDD approach is a better choice for both WF and MSIG-fix1 approaches.

While WF and MSIG-fix1 are the best gain functions to be used for the MDD approach in terms of the KurtR measure, LSA performs the best for the DD approach with

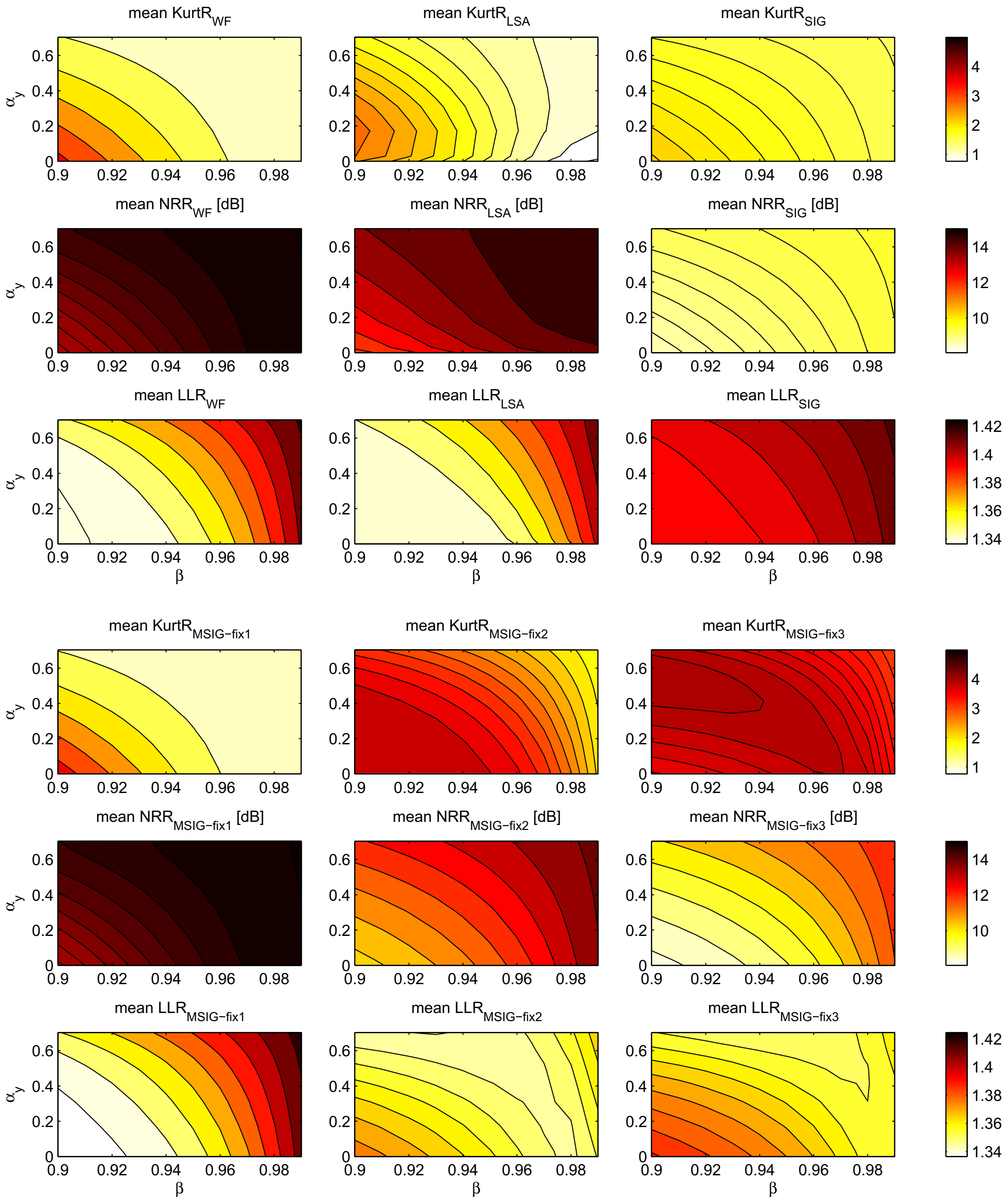

Fig. 11. Average results with $\hat{\xi}_{DD}$ at SNR = 0 dB.

large smoothing parameters. On the other hand, for less smoothing applied, SIG is the best gain functions among all. Apart from that, MSIG-fix2 and MSIG-fix3 perform worst in KurtR results among all evaluated gain functions. This is because both gain functions are less aggressive and are not directly propositional towards the *a priori* SNR as shown in Fig. 1. There is drop in gain values between $\xi = -5$ dB and $\xi = 0$ dB, which should be the main factor for the isolated spectral components to be formed after the processing.

The results from the NRR measure are almost inversely proportional to the KurtR measure. All evaluated gain functions show poorer performance for the MDD approach when compared to the DD approach. This can be explained as MDD has more variations in noise when compared to the DD approach. However, the remaining residual noise in the enhanced signal helps to mask the musical noise. This acts as a good compromise between the amount of musical noise and noise reduction for the MDD approach.

In terms of the amount of speech distortion generated, LLR is almost directly proportional to the NRR measure. From the figures, a small LLR can be obtained for a decreasing β and α_y. At small β and α_y values, the DD

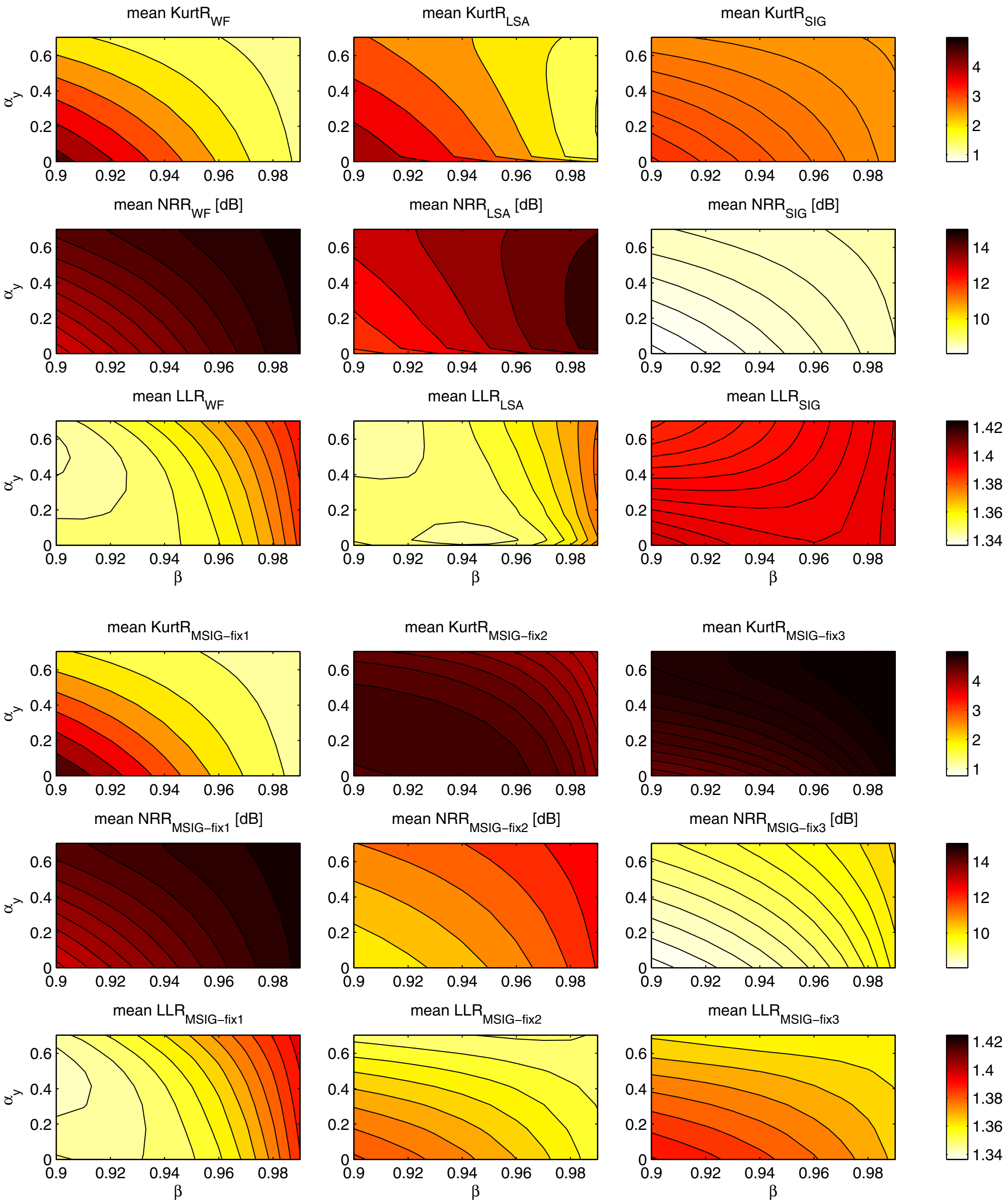

Fig. 12. Average results with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 0 dB.

approach performs better than the MDD approach. Meanwhile, at large β and α_y values, the MDD approach generally gives less speech distortion when compared to the DD approach. This indicates that with a large smoothing parameters, the MDD approach performs better than the DD approach, particularly for the WF, LSA and MSIG-fix1 gain functions.

6.5. Subjective listening tests

Subjective listening tests were performed to validate the results from the objective measures. The description of the tests setup can be found in Section 6.2. Prior to the subjective tests, appropriate choice of parameters β and α_y has to be found.

6.5.1. Selection of smoothing parameters

Tables 3 and 4 give the best smoothing parameters for different gain functions and different SNR levels. As can be seen in the tables, a small value for β is preferred for lower speech distortion in the case of WF, LSA, SIG and MSIG-fix1 gain functions. This is validated from the LLR, PESQ and SNRseg results. However, there is a trade-off because of the resulting high level of annoying musical noise in the output when both β and α_y are small. Also, small values in β and α_y give low NRR values, which

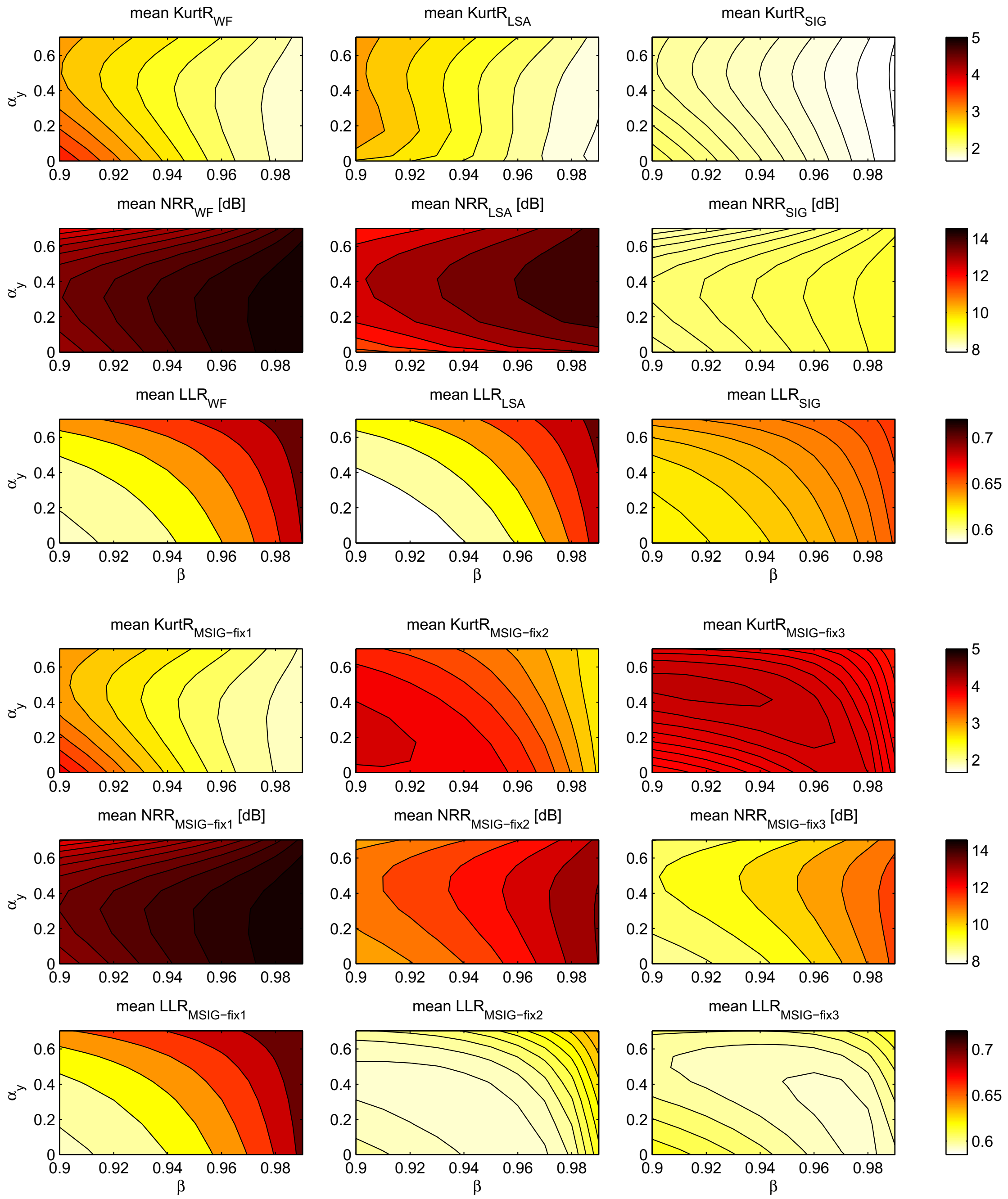

Fig. 13. Average results with $\hat{\xi}_{DD}$ at SNR = 15 dB.

is not desirable in most situations. In addition, an increment of both β and α_y does not give a direct drop in speech quality. This means that by choosing appropriate values for both smoothing parameters, the balanced trade-off can be obtained. Meanwhile, for MSIG-fix2 and MSIG-fix3, better PESQ, SNRseg and LLR results are recorded at larger values for β and α_y. This motivates the use of MSIG functions as they are able to show similar preferred smoothing parameters from all the objective measurement results.

The last columns in Tables 3 and 4 also show the smoothing parameters chosen for the listening test. The value β for each gain function and each SNR condition was chosen such that the smallest value was selected so that the KurtR was at its minimum value. This is to keep the speech distortion as low as possible while having the lowest possible level of audible musical noise. The value α_y were chosen as the mean of α_y values from all the objective measures. This have been found to have a reasonable compromise between the aforementioned trade-offs.

6.5.2. *Objective measurement with selected smoothing parameters*

Objective measurement was also performed for the selected parameters listed in Tables 3 and 4, using the NOIZEUS database. Tables 5 lists the average results for pink noise, where Δ indicates the difference between the

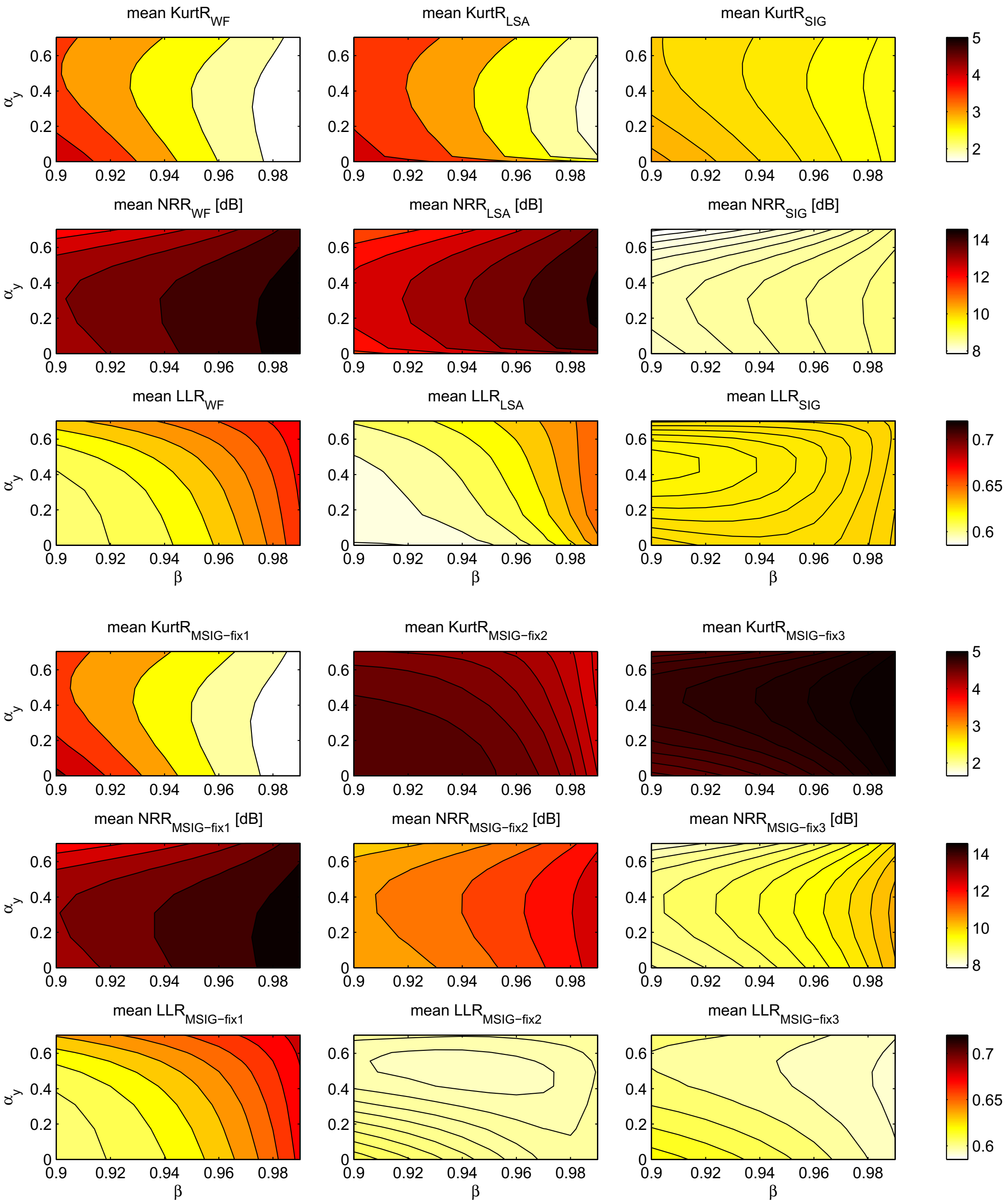

Fig. 14. Average results with $\hat{\xi}_{\mathrm{MDD}}$ at SNR = 15 dB.

enhanced signals and the noisy signals. According to the table, WF, LSA, and MSIG-fix1 have better performance particularly in terms of noise reduction as reflected by the results from SNRseg and NRR measures. When comparing both DD and MDD *a priori* SNR estimates, WF and MSIG-fix1 perform best with MDD approach, as indicated by larger NRR results while having smaller KurtR values. While for LSA, DD has slightly less musical noise when compared to MDD, with approximately similar NRR results. For MSIG-fix2 and MSIG-fix3, their results indicate less noise reduction when compared to WF, LSA and MSIG-fix1. In terms KurtR measure, MSIG-fix2 with DD approach has similar amount of musical noise when compared to WF and MSIG-fix1. For MSIG-fix2 with MDD approach and MSIG-fix3 with both DD and MDD approaches, KurtR results are large. In terms of the least amount of musical noise generated, the SIG function has the smallest KurtR values, but with the least amount of noise reduction from SNRseg and NRR measures. As for the amount of speech distortion, all evaluated gain functions have almost similar performance, as shown by PESQ and LLR scores, except for SIG with slightly poorer performance when compared to others.

Table 6 shows the average results for the factory noise. The results are similar to the results from pink noise, except for some cases from the KurtR values. As observed in

Table 3
Approximated best smoothing values for different gain functions with DD approach.

		SNR	PESQ	SNRseg	KurtR	NRR	LLR	Test
WF	β	0 dB	0.90	0.90	0.91	0.94	0.90	0.91
		15 dB	0.90	0.90	0.98	0.95	0.90	
	α_y	0 dB	0–0.2	0–0.4	0.7	0.7	0–0.3	0.35
		15 dB	0	0–0.2	0–0.5	0.2–0.3	0–0.2	
LSA	β	0 dB	0.90	0.90	0.97	0.95	0.90	0.95
		15 dB	0.90	0.90	0.99	0.96	0.90	
	α_y	0 dB	0.1–0.5	0.3–0.4	0	0.6–0.7	0–0.7	0.39
		15 dB	0.1–0.4	0.2	0–0.2	0.3–0.4	0–0.4	
SIG	β	0 dB	0.90	0.90	0.99	0.98	0.90	0.98
		15 dB	0.90	0.90	0.99	0.98	0.90	
	α_y	0 dB	0–0.1	0–0.1	0.6	0.6–0.7	0–0.6	0.40
		15 dB	0–0.3	0–0.2	0.2–0.6	0–0.5	0–0.3	
MSIG-fix1	β	0 dB	0.90	0.90	0.90	0.94	0.90	0.90
		15 dB	0.90	0.90	0.98	0.97	0.90	
	α_y	0 dB	0–0.2	0–0.4	0.7	0.6–0.7	0–0.4	0.35
		15 dB	0	0–0.2	0–0.5	0.1–0.3	0–0.1	
MSIG-fix2	β	0 dB	0.90	0.93	0.98	0.99	0.90	0.98
		15 dB	0.94	0.90	0.98	0.99	0.90	
	α_y	0 dB	0.6–0.7	0.5	0.7	0.2–0.7	0.6–0.7	0.52
		15 dB	0.2–0.3	0–0.5	0.7	0–0.4	0	
MSIG-fix3	β	0 dB	0.95	0.96	0.99	0.97	0.90	0.97
		15 dB	0.95	0.94	0.99	0.99	0.90	
	α_y	0 dB	0.5–0.6	0.5–0.6	0.6–0.7	0.6–0.7	0.7	0.52
		15 dB	0.3	0.2–0.3	0.6–0.7	0–0.6	0	

Table 4
Approximated best smoothing values for different gain functions with MDD approach.

		SNR	PESQ	SNRseg	KurtR	NRR	LLR	Test
WF	β	0 dB	0.90	0.90	0.96	0.98	0.90	0.96
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0–0.6	0.2–0.4	0.7	0.6	0.4-0.6	0.49
		15 dB	0–0.3	0–0.3	0–0.6	0–0.4	0–0.4	
LSA	β	0 dB	0.90	0.94	0.97	0.98	0.90	0.97
		15 dB	0.90	0.90	0.98	0.99	0.90	
	α_y	0 dB	0.3–0.6	0.2–0.3	0.5	0.2–0.4	0.4–0.6	0.36
		15 dB	0.2	0–0.2	0.2–0.4	0.2–0.4	0	
SIG	β	0 dB	0.90	0.97	0.98	0.97	0.90	0.97
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0.6	0.2-0.4	0.7	0.6	0.7	0.49
		15 dB	0.2	0.1–0.2	0.2–0.6	0.1–0.4	0.4–0.5	
MSIG-fix1	β	0 dB	0.90	0.90	0.96	0.97	0.90	0.96
		15 dB	0.90	0.90	0.98	0.98	0.90	
	α_y	0 dB	0.2–0.4	0–0.5	0.6	0.7	0.3–0.6	0.48
		15 dB	0–0.3	0–0.3	0–0.6	0–0.4	0–0.4	
MSIG-fix2	β	0 dB	0.93	0.99	0.99	0.97	0.94	0.97
		15 dB	0.98	0.90	0.99	0.98	0.91	
	α_y	0 dB	0.7	0.2-0.7	0.7	0.7	0.7	0.56
		15 dB	0.2–0.3	0–0.2	0.4–0.7	0.3	0.5–0.6	
MSIG-fix3	β	0 dB	0.97	0.99	0.90	0.99	0.90	0.90
		15 dB	0.99	0.98	0.90	0.99	0.99	
	α_y	0 dB	0.6–0.7	0.4–0.7	0	0.7	0.7	0.50
		15 dB	0.1–0.3	0.2	0.7	0.1–0.4	0.1–0.6	

15 dB SNR, the DD approach has smaller KurtR results for all evaluated gain functions when compared to the MDD approach for factory noise. Here, MSIG-fix3 has the smallest KurtR results, which is similar to the results from the SIG function. However, since MSIG-fix3 has larger SNRseg and similarly small NRR results when compared with the SIG function, more distorted noise would be perceived as musical noise.

Table 5
Objective results for pink noise with selected parameters.

Signal	ΔPESQ		ΔSNRseg		KurtR		NRR		ΔLLR	
	DD	MDD	DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB SNR										
WF	0.5	0.5	3.9	3.9	2.3	1.8	14.2	14.6	−0.2	−0.2
LSA	0.5	0.5	3.7	3.9	1.5	2.1	14.4	14.3	−0.2	−0.2
SIG	0.3	0.4	2.7	2.9	1.6	2.6	9.5	8.7	−0.1	−0.1
MSIG-fix1	0.5	0.5	4.0	3.9	2.5	1.7	14.1	14.6	−0.2	−0.2
MSIG-fix2	0.5	0.5	3.6	3.7	2.2	4.3	13.8	12.3	−0.2	−0.2
MSIG-fix3	0.5	0.4	3.4	3.1	3.7	4.8	11.5	9.0	−0.2	−0.2
15 dB SNR										
WF	0.6	0.6	3.5	3.4	2.9	2.3	13.7	14.1	−0.3	−0.3
LSA	0.6	0.6	3.1	3.5	2.3	2.3	13.9	14.1	−0.3	−0.3
SIG	0.5	0.5	3.0	3.4	1.7	2.5	9.3	8.7	−0.2	−0.3
MSIG-fix1	0.6	0.6	3.5	3.4	3.2	2.3	13.6	14.1	−0.3	−0.2
MSIG-fix2	0.6	0.6	2.9	3.6	2.9	4.3	13.1	12.0	−0.3	−0.3
MSIG-fix3	0.5	0.5	3.3	3.4	4.0	4.8	10.9	8.9	−0.3	−0.3

Table 6
Objective results for factory noise with selected parameters.

Signal	ΔPESQ		ΔSNRseg		KurtR		NRR		ΔLLR	
	DD	MDD	DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB SNR										
WF	0.6	0.6	4.6	4.5	4.1	4.1	13.9	14.3	−0.1	−0.1
LSA	0.6	0.6	4.5	4.6	3.4	4.0	14.2	14.1	−0.1	−0.1
SIG	0.4	0.4	3.4	3.4	2.2	3.1	9.8	9.0	−0.1	−0.2
MSIG-fix1	0.6	0.6	4.6	4.5	4.2	4.1	13.8	14.4	−0.1	−0.1
MSIG-fix2	0.6	0.6	4.4	4.5	3.4	4.1	13.4	12.2	−0.1	−0.2
MSIG-fix3	0.5	0.5	4.3	4.0	3.4	3.9	11.4	9.3	−0.2	−0.2
15 dB SNR										
WF	0.5	0.5	5.0	4.9	3.8	4.2	12.3	12.5	0.1	0.1
LSA	0.5	0.6	4.7	4.9	3.6	4.3	12.4	13.0	0.1	0.1
SIG	0.4	0.4	4.3	4.4	2.4	2.9	9.1	8.5	0	0
MSIG-fix1	0.5	0.5	5.0	4.9	3.8	4.3	12.1	12.5	0.1	0.1
MSIG-fix2	0.5	0.5	4.6	5.1	3.3	3.4	11.6	10.7	0.1	0
MSIG-fix3	0.4	0.4	5.0	5.0	2.8	2.9	9.8	8.3	0	0

Table 7
Subjective results for pink noise.

SNR	Signal	SPCH		NSE		MUSIC		Overall	
		DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB	Noisy	3.0		1.0		5.0		2.98	
	WF	3.5	3.6	2.6	2.8	3.4	3.5	3.16	3.28
	LSA	3.3	3.5	2.2	2.4	4.1	3.8	3.21	3.22
	SIG	3.0	3.2	1.4	1.5	4.6	4.1	2.97	2.92
	MSIG-fix1	3.7	3.7	2.9	2.9	3.1	3.5	3.20	3.35
	MSIG-fix2	3.8	4.0	3.0	3.7	2.5	1.7	3.08	3.12
	MSIG-fix3	4.1	4.2	4.0	4.2	1.3	1.0	3.12	3.12
15 dB	Noisy	4.0		1.6		5.0		3.51	
	WF	4.6	4.6	3.5	3.6	3.7	4.0	3.91	4.09
	LSA	4.4	4.4	3.1	3.3	4.2	4.1	3.91	3.94
	SIG	4.1	4.2	2.1	2.4	4.8	4.1	3.66	3.54
	MSIG-fix1	4.7	4.9	3.8	3.7	3.6	4.0	3.99	4.18
	MSIG-fix2	4.9	4.9	3.8	4.3	3.4	2.5	4.00	3.86
	MSIG-fix3	4.9	4.9	4.4	4.5	2.4	1.5	3.86	3.58

Table 8
Subjective results for factory noise.

SNR	Signal	SPCH		NSE		MUSIC		Overall	
		DD	MDD	DD	MDD	DD	MDD	DD	MDD
0 dB	Noisy	3.9		1.7		5.0		3.52	
	WF	4.6	4.6	3.5	3.6	4.5	4.6	4.18	4.26
	LSA	4.4	4.5	3.3	3.3	4.7	4.6	4.14	4.14
	SIG	4.1	4.2	2.4	2.4	4.9	4.6	3.79	3.73
	MSIG-fix1	4.7	4.7	3.6	3.6	4.3	4.5	4.18	4.28
	MSIG-fix2	4.7	4.7	3.7	3.9	4.0	3.1	4.13	3.90
	MSIG-fix3	4.7	4.7	3.9	4.0	2.6	2.1	3.75	3.62
15 dB	Noisy	4.5		2.8		5.0		4.09	
	WF	4.9	4.9	4.1	4.1	4.9	4.9	4.61	4.62
	LSA	4.9	4.9	4.0	4.1	4.9	4.9	4.58	4.59
	SIG	4.7	4.7	3.4	3.4	4.9	4.7	4.33	4.29
	MSIG-fix1	4.9	4.9	4.1	4.1	4.8	4.9	4.60	4.63
	MSIG-fix2	4.9	4.9	4.2	4.2	4.8	4.3	4.60	4.45
	MSIG-fix3	4.9	4.9	4.2	4.3	4.1	3.7	4.40	4.27

6.5.3. Evaluation for listening tests

Tables 7 and 8 tabulate the average results of the subjective listening tests, and also the overall performance of each gain function by taking the average of SPCH, NSE and MUSIC. From the overall scales, it can be observed that although the difference is not that significant, the listeners preferred the signals with WF and MSIG-fix1 gain functions, both combined with the MDD approach. The results are consistent for both noise types. For the LSA method, the overall performance between the DD and the MDD approach are almost identical, while LSA with DD approach has recorded the least amount of musical noise particularly in the pink noise. For the MSIG-fix2 and MSIG-fix3 functions the audible musical noise is more prominent which is reflected in the MUSIC column. This trend is particularly obvious with the MDD approach, which aligns with the objective results. As for the SIG function there is less musical noise in the output, but also a very small amount of noise suppression. The MDD approach helps to reduce audible noise and increase noise reduction for aggressive gain functions such as WF and MSIG-fix1. While for other less aggressive gain functions, the DD approach can be sufficient. As for factory noise, particularly at 15 dB SNR, the subjective results are almost the same for each *a priori* SNR estimate and every gain function. The suggested reason behind this is that the factory noise is less intrusive than the pink noise. Thus, less noise with less audible musical noise was perceived when the noise floor was fixed at -15 dB for the gain functions.

7. Conclusions

In this paper, a new MSIG has been developed to provide flexibility to the gain function that can be optimized to match various criteria to achieve a compromised trade-off between speech distortion, noise reduction and musical noise. In addition, a new approach to estimate the *a priori* SNR has been proposed for the MDD, which reduces and eliminates the speech transient distortion. The musical noise has been further reduced by applying more smoothing to the *a posteriori* SNR by using a recursive averaging algorithm. As such, the level of smoothing is controlled by the parameters β and α_y. Performance evaluation shows that the proposed MDD performs better than the traditional DD. The MSIG-fix1 function has similar performance compared to the conventional gain functions. At large smoothing parameters, MSIG-fix2 and MSIG-fix3 generate the lowest speech distortion among all evaluated gain functions. Finally, subjective listening tests verify the findings from the objective measurements and give confidence that the chosen objective measures reflect the true subjective experience.

Future work includes combining the MSIG functions by incorporating more prior information of speech, such as the pitch information and perceptual information in the design of the parameters. The objective measures can be used to optimize the MSIG gain function with respect to speech quality and intelligibility.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2012.09.004.

References

Alam, M., Chowdhury, M., Alam, M., 2009. Comparative study of a priori signal-to-noise ratio (SNR) estimation approaches for speech enhancement. J. Electr. Electron. Eng. 9 (1), 809–817.

Andrianakis, I., White, P., 2009. Speech spectral amplitude estimators using optimally shaped gamma and chi priors. Speech Comm. 51 (1), 1–14.

Berouti M., Schwartz R., Makhoul J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'79), vol. 4, pp. 208–211.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process 27 (2), 113–120.

Breithaupt, C., Martin, R., 2011. Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions. IEEE Trans. Audio Speech Lang. Process. 19 (2), 277–289.

Breithaupt, C., Gerkmann, T., Martin, R., 2008. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In: Proc. IEEE Internat. Conf. on Acoustics Speech, and Signal Processing (ICASSP'08), pp. 4897–4900.

Breithaupt, C., Krawczyk, M., Martin, R., 2008. Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'08), pp. 4037–4040.

Cappé, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2 (2), 345–349.

Chang, J., Kim, N., Mitra, S., 2006. Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. 54 (6), 1965–1976.

Cohen, I., 2004. Speech enhancement using a noncausal a priori SNR estimator. IEEE Signal Process. Lett. 11 (9), 725–728.

Davis, A., Nordholm, S., Low, S.Y., Togneri, R., 2006. A multi-decision sub-band voice activity detector. In: Proc. 14th European Signal Processing Conf. (EUSIPCO'06), Florence, Italy.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Gustafsson, S., Martin, R., Jax, P., Vary, P., 2002. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. IEEE Trans. Speech Audio Process. 10 (5), 245–256.

Hansen, J., Pellom, B., 1998. An effective quality evaluation protocol for speech enhancement algorithms. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 2819–2822.

Hendriks, R., Heusdens, R., Jensen, J., 2010. MMSE based noise PSD tracking with low complexity. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'10), pp. 4266 – 4269.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238.

Loizou, P., 2007. Speech Enhancement Theory and Practice. CRC Press, Boca Raton, FL.

Paliwal, K., Wójcicki, K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm. 52 (5), 450–475.

Paliwal, K., Schwerin, B., Wójcicki, K., 2012. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Comm. 54 (2), 282–305.

Park, Y.S., Chang, J.H., 2007. A novel approach to a robust a priori SNR estimator in speech enhancement. IEICE Trans. Comm. E90-B (8), 2182–2185.

Plapous, C., Marro, C., Scalart, P., 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14 (6), 2098–2108.

Plourde, E., Champagne, B., 2009. Generalized bayesian estimators of the spectral amplitude for speech enhancement. IEEE Signal Process. Lett. 16 (6), 485–488.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs, NJ.

Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ), a new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'01), vol. 2, pp. 749–752.

Scalart, P., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96), vol. 2, pp. 629–632.

Suhadi, S., Last, C., Fingscheidt, T., 2011. A data-driven approach to a priori SNR estimation. IEEE Trans. Audio Speech Lang. Process. 19 (1), 186–195.

Uemura, Y., Takahashi, Y., Saruwatari, H., Shikano, K., Kondo, K., 2008. Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics. In: Proc. Internat. Workshop on Acoustic Echo and Noise Control (IWAENC'08), Seattle, USA.

Yong, P.C., Nordholm, S., Dam, H.H., Low, S.Y., 2011. On the optimization of sigmoid function for speech enhancement. In: Proc. 19th Eur. Signal Process. Conf. (EUSIPCO'11), Barcelona, Spain, pp. 211–215.