Circuits, Systems, and Signal Processing (2021) 40:2926–2949
https://doi.org/10.1007/s00034-020-01604-6

DNN-Based Calibrated-Filter Models for Speech Enhancement

Yazid Attabi[1] · Benoit Champagne[1] · Wei-Ping Zhu[2]

Received: 23 April 2020 / Revised: 10 November 2020 / Accepted: 13 November 2020 /
Published online: 27 January 2021

Abstract

In this paper, we present a new two-stage speech enhancement approach, specially conceived to reduce musical and other random noises without requiring their localization in the time–frequency domain.The proposed method is motivated by two observations: (1) the random scattering nature of the energy peaks corresponding to the musical noise in the spectrogram of the processed speech; and (2) the existence of correlation between Wiener filter gains calculated at different frequencies. In the first stage of the proposed method, a preliminary gain function is generated using the nonnegative matrix factorization algorithm. In the second stage, a modified gain function that is more robust to noise artefacts, and referred to as *calibrated filter*, is estimated by applying a DNN-based nonlinear mapping function to the preliminary gain function. To further decrease the variability of the estimated calibrated filter, we propose to expand the DNN-based extraction of frequency dependencies to a set of preliminary gain functions derived from spectral estimates based on a family of data tapers; the resulting calibrated filter is referred to as *multi-filter*. The evaluation of the proposed DNN-based calibrated filter models for speech enhancement, under different noise types and input SNR levels, shows substantial improvements in terms of standard speech quality and intelligibility measures when compared to uncalibrated filter.

Keywords Speech enhancement · Wiener filter gain function · Gain calibration · Multi-taper spectral analysis · Deep neural network (DNN) · Non-negative matrix factorization (NMF)

✉ Yazid Attabi
yazid.attabi@mcgill.ca

Extended author information available on the last page of the article

1 Introduction

Speech enhancement aims to improve the quality and intelligibility of speech by isolation of the target speech from contaminating background noises. Several algorithms for speech enhancement involving a single audio channel have been proposed in the past, e.g., [19, 37, 48, 53]. Most single-channel speech enhancement methods decompose the audio signal of the noisy speech in the frequency domain and weight the spectral coefficients using an estimated gain function. The latter provides the amount of attenuation (or gain) that must be applied to the noisy speech spectrum at any given frequency to obtain the enhanced speech spectrum. The suppression function, Wiener filter and ideal ratio mask (IRM) are examples of such gain functions. The suppression function is used in the spectral-subtractive algorithms, in which the clean speech spectrum is estimated by subtracting the noise spectrum from the noisy speech spectrum [3, 4, 24, 44]. The Wiener filter is derived from the optimization of a linear time-invariant filter, aiming to minimize the mean square error between the desired signal and its estimate [5, 20, 31, 38, 40, 52]. IRM is used as a target value of a complex nonlinear mapping function estimated using deep neural networks (DNN) [11, 35, 50, 51].

Each one of these speech enhancement methods produces a distinct shape of the gain function, which can provide an improvement in terms of noise attenuation. As a side effect, the gain function also alters the quality of the original clean speech. Therefore, it is important to find a proper balance between the amounts of noise reduction and introduced speech distortion [32]. Indeed, a poor estimation of the enhancement gain introduces isolated spectral energy peaks of short duration at random positions in the processed audio spectrum. These isolated components, known as musical noise artifacts, are perceived as unpleasant tones that lead to a serious deterioration of the speech quality, particularly under low signal-to-noise ratio (SNR) conditions and during speech pauses [12]. The main factors responsible for musical noise include [32]: (1) nonlinear processing of the power spectrum, (2) inaccurate estimation of the noise spectrum, (3) large variance in the estimates of the noise and noisy speech signal spectra, and (4) large variability in the gain function.

Several methods have been proposed in the past to eliminate the musical noise. Most of these methods were designed to separately tackle one of the above contributing factors. Regarding the first factor (i.e., nonlinear processing), the iterative spectral subtraction has been proposed in [26, 30, 33, 54]. This method assumes that a weak nonlinear processing can reduce musical noise generation when iteratively applied to the input signal. Under the assumption of stationary input noise, it can improve speech quality with low musical noise. To address the second factor (inaccurate noise estimation), some researchers have developed more efficient speech pause detectors, which play a crucial role in noise spectrum estimation [12, 14, 41, 55]. In [12], the authors also proposed a postfilter as a second step to further reduce musical noise. The postfilter, which adaptively smooths the gain function over frequency based on soft-decisions from a low-SNR detector, leads to consistent improvements of speech quality. The third factor (large spectral variance) can be addressed through modifications of the classical windowed periodogram-type estimator. To this end, the use of multi-tapering along with wavelet thresholding has been proposed in [20, 43]. In this method, a set of orthogonal tapers is applied to the speech signal and the resulting spectral estimates are then averaged, which reduces the spectral variance. Recently, multi-tapering has been employed to enhance the speech/noise dictionary and activation matrices in the nonnegative matrix factorization (NMF) method [1]. Regarding the fourth factor (gain function variability), the adaptive exponential time-averaging method was proposed to smooth either the Wiener gain function [17] or the a priori SNR needed in its calculation. [10, 40].

Some researchers have taken another approach to reduce the musical noise, which consists in the localization and elimination of the isolated peaks in the spectrogram of the enhanced speech [2, 3, 15]. The peak localization also serves to assess the amount of musical noise present in the enhanced speech, creating an objective measure of musical noise [9, 18]. Determining the presence of isolated spectrogram peaks involves several processing steps. In [18], this includes detection of small local minima in the spectrogram, application of Delaunay triangulation over local minima, selection of specific triangles, and grouping of adjacent triangles in domains. The method in [9] involves the following steps: detection of isolated peaks, verification of the non-harmonicity condition, and detection of transient spectral components. In the context of musical noise reduction, the peak localization process remains a complex and error-prone task.

In this work, we propose a supervised machine learning-based speech enhancement approach in two stages, specially designed to reduce musical noises from processed speech, without requiring the localization of isolated noise peaks in the time–frequency domain. The underlying idea is based on two key observations: (1) the random scattering nature of the energy peaks corresponding to the musical noise in the spectrogram, and (2) the existence of dependencies across frequency bins in the gain functions used for speech enhancement. Specifically, in the first stage of our proposed approach, a preliminary gain function is estimated using a robust speech enhancement algorithm, which in this work is based on the NMF method [49]. In the second stage, the accuracy of the estimated gain function is refined using a DNN-based nonlinear mapping. We refer to the impact of this process on the gain function as a *calibration effect*. Subsequently, instead of using a single gain function, we extend the proposed method to a set of gain functions with different spectral properties, which will be combined using a DNN-based fusion approach. By using diverse gain functions, we can decrease the variability in the estimated filter, and hence, further enhance the quality of the processed speech. We refer to the impact of using a set of gain functions in this manner as a *fusion effect*.

The paper is organized as follows. Section 2 introduces the proposed DNN-based calibrated filter model for speech enhancement. Section 3 explains how the concept of a calibrated gain function can be extended to a set of multiple gain functions. Section 4 reports on the experimental speech enhancement performance of the proposed systems using different objective measures. Finally, Section 5 concludes the paper.

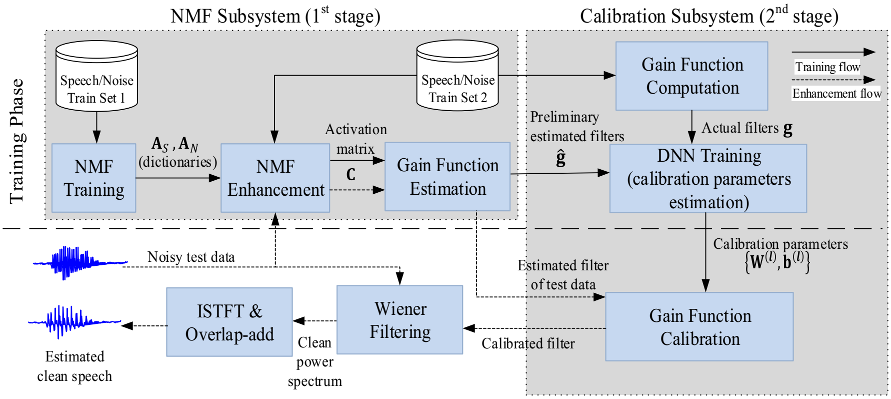
NMF Subsystem (1st stage)
Calibration Subsystem (2nd stage)
Training Phase
Speech/Noise Train Set 1
Speech/Noise Train Set 2
Training flow
Enhancement flow
Gain Function Computation
NMF Training
$\mathbf{A}_S, \mathbf{A}_N$ (dictionaries)
NMF Enhancement
Activation matrix
$\mathbf{C}$
Gain Function Estimation
Preliminary estimated filters
$\hat{\mathbf{g}}$
Actual filters $\mathbf{g}$
DNN Training (calibration parameters estimation)
Calibration parameters
$\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}$
Noisy test data
Estimated filter of test data
ISTFT & Overlap-add
Wiener Filtering
Gain Function Calibration
Clean power spectrum
Calibrated filter
Estimated clean speech

Fig. 1 Block diagram of the proposed calibrated gain function-based speech enhancement system

2 DNN-Based Calibrated Filter Model

The proposed speech enhancement system using calibrated gain function consists of two main components, as shown in Fig. 1, that is: an NMF sub-system, used for the first stage, and a calibration sub-system, used for the second stage. The NMF sub-system estimates the power spectra of the clean speech and the noise and generates the preliminary gain function. The calibration sub-system refines the accuracy of the gain function estimated in the first stage using a DNN-based nonlinear mapping. In this work, we chose the NMF algorithm in the first stage due to its capability to recover clean speech from noisy observations without relying on the stationarity assumption for the additive noise [8, 25, 28]. However, the first stage of the proposed framework can support other enhancement algorithms that produce a gain function.

2.1 NMF-Based Speech Enhancement Sub-system

In single-channel speech enhancement, the time-domain noisy speech signal $y(t)$ is composed of the clean speech signal $s(t)$ and the additive noise signal $n(t)$, that is,

$$y(t) = s(t) + n(t) \tag{1}$$

where t is the discrete-time sample index. The noisy speech spectrum, obtained via short-time Fourier transform (STFT) of consecutive overlapping frames, can be expressed as $Y_{kj} = S_{kj} + N_{kj}$, where j represents the frame index, $k = 0, \ldots, K - 1$, is the frequency bin index, $K = F/2$, and F is the frame size.[1] In NMF-based speech enhancement, we assume in practice that the magnitude spectrum of the noisy speech, obtained via STFT, can be approximated by the sum of the clean speech and

[1] Only half of the coefficients are used since the audio signal samples are real-valued and their spectral coefficients exhibit complex conjugate symmetry.

noise magnitude spectra, i.e., $\left|Y_{kj}\right|^{\nu} \approx \left|S_{kj}\right|^{\nu} + \left|N_{kj}\right|^{\nu}$ with $\nu = 1$ being the most common choice [34, 49].

In speech and audio applications, NMF interprets the magnitude or power spectrum of the target signal as a linear combination of basis vectors, which play a key role in the enhancement or separation process. Specifically, NMF decomposes a given matrix into a product of a basis (or dictionary) matrix and an activation (or encoding) matrix with non-negative elements constraint [13, 29]. For a nonnegative matrix $\mathbf{V} = \left[v_{kj}\right] \in \mathbb{R}_{+}^{K\times J}$, NMF aims to find a local optimal decomposition of $\mathbf{V} = \mathbf{AC}$, where $\mathbf{A} = \left[a_{km}\right] \in \mathbb{R}_{+}^{K\times M}$ is a basis matrix, $\mathbf{C} = \left[c_{mj}\right] \in \mathbb{R}_{+}^{M\times J}$ is an activation matrix, $\mathbb{R}_{+}$ denotes the set of nonnegative real numbers, M is the number of basis vectors, and J is the number of consecutive frames. The factorization is obtained by minimizing the reconstruction error between the observation matrix $\mathbf{V}$ and the model $\mathbf{AC}$ using the Kullback–Leibler (KL) divergence as a cost function, while constraining the matrices to be entry-wise nonnegative. The solutions can be obtained iteratively using the following multiplicative update rules [29],

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{(\mathbf{V}/\mathbf{AC})\mathbf{C}^{\mathrm{T}}}{\mathbf{1}\mathbf{C}^{\mathrm{T}}}, \quad \mathbf{C} \leftarrow \mathbf{C} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{V}/\mathbf{AC})}{\mathbf{A}^{\mathrm{T}}\mathbf{1}} \tag{2}$$

where the operation $\otimes$ denotes element-wise multiplication, / and the quotient line is element-wise division, $\mathbf{1}$ is a $K \times J$ matrix with ones, and the superscript T is the matrix transpose. In this work, $\mathbf{V} = [v_{kj}]$ contains the magnitude spectrum values of either one of the noisy speech, clean speech, and noise, as indicated by subscripts or superscripts Y, S, and N, respectively.

In a supervised framework, the $\mathbf{A}$ matrices of clean speech and noise, denoted as $\mathbf{A}_S$ and $\mathbf{A}_N$, respectively, are first obtained during the training stage, by applying both update rules in (2) to the training data $\mathbf{V}_S$ and $\mathbf{V}_N$. In the enhancement stage, the activation matrix $\mathbf{C}_Y = \left[\mathbf{C}_S^{\mathrm{T}}\mathbf{C}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ is estimated by applying only the activation update to $\mathbf{V}_Y$, while fixing the basis matrix $\mathbf{A}_Y = \left[\mathbf{A}_S\mathbf{A}_N\right]$. Then, the clean speech spectrum can be estimated using a Wiener filter as [13, 28],

$$\widehat{S}_{kj} = \widehat{g}_{kj} Y_{kj} \tag{3}$$

where $\widehat{g}_{kj}$ is the estimated Wiener gain function, given by,

$$\widehat{g}_{kj} = \frac{\widehat{P}_{kj}^{S}}{\widehat{P}_{kj}^{S} + \widehat{P}_{kj}^{N}} \tag{4}$$

In this expression, $\widehat{P}_{kj}^{S}$ and $\widehat{P}_{kj}^{N}$ denote the estimated power spectra of the clean speech and noise, respectively. The latter are obtained via temporal smoothing of the NMF-based periodograms as [28],

$$\begin{aligned} \widehat{P}_{kj}^{S} &= \tau_{\mathrm{s}} \widehat{P}_{k(j-1)}^{S} + \left(1 - \tau_{\mathrm{s}}\right)\left[\mathbf{A}_S\mathbf{C}_S\right]_{kj}^{2} \\ \widehat{P}_{kj}^{N} &= \tau_{\mathrm{N}} \widehat{P}_{k(j-1)}^{N} + \left(1 - \tau_{\mathrm{N}}\right)\left[\mathbf{A}_N\mathbf{C}_N\right]_{kj}^{2} \end{aligned} \tag{5}$$

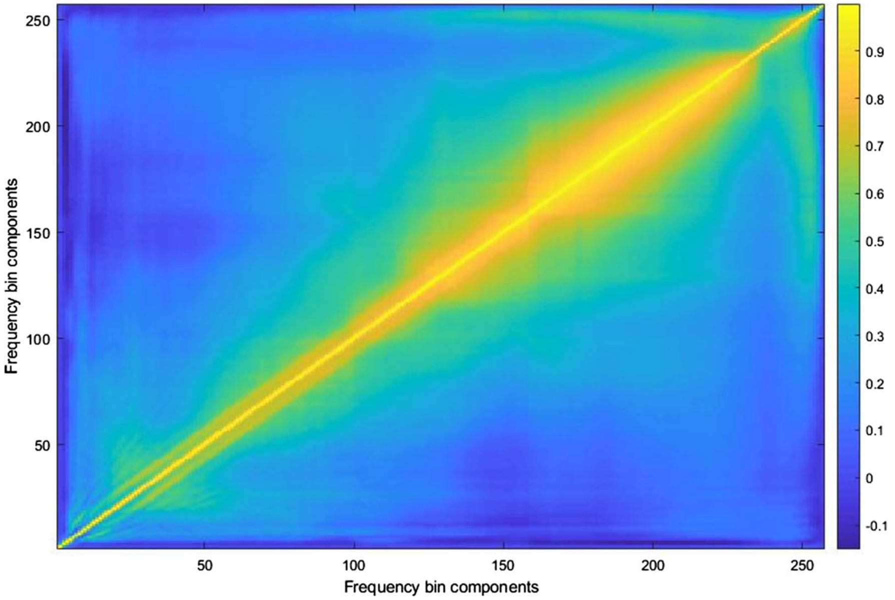

Fig. 2 Plot of the sample correlation matrix of actual Wiener filter gain vector, computed using clean speech and different types of noise from the training data (color level indicates normalized value of corresponding matrix entry)

where τ_{s} and τ_{N} are the smoothing factors for the speech and noise, and $[\bullet]_{kj}$ denotes the (k,j)-th entry of its matrix argument. In the sequel, we shall refer to the gain function $\widehat{g}_{kj}$ (4) as the preliminary (or unprocessed) gain function.

2.2 DNN-Based Calibration Sub-system

The goal of the calibration process is to reduce the musical or residual noise, without the complex task of localizing or assessing the isolated noise spectral peaks. The idea underlying the calibration method is based on two observations. First, and by definition, the isolated spectral peaks of energy leading to audible musical noise are located at random positions (i.e., time and frequency) in the spectrogram. An estimated spectral component $\widehat{S}_{kj}$ corresponding to such an isolated noisy peak at frequency bin k and frame j is usually surrounded by spectral values with much smaller magnitudes. Subsequently, the filter weight values in the immediate neighborhood of bin k, i.e., $\widehat{g}_{qj}$ for $q \neq k$, should also be relatively small.

The second observation is the existence, for noisy speech, of a relationship between the values of gain function calculated at nearby frequencies. Such a dependency is illustrated in Fig. 2, which displays the sample *Pearson* correlation matrix for the *actual* Wiener filter gain vector $\mathbf{g}_j = \left[g_{1j}g_{2j}\cdots g_{Kj}\right]^{\mathrm{T}}$, computed using clean speech and noise from a training data set (as per the methodology described in Sect. 4). We can observe a strong level of correlation between values of g_{kj}

computed at neighboring frequency bins, suggesting the existence of a relationship between these components. Furthermore, the range of the dependent neighbors is particularly important for the mid to high frequency band. When an estimated spectral component $\widehat{S}_{kj}$, resulting from an inaccurate gain estimate $\widehat{g}_{kj}$, corresponds to musical noise, it can be corrected $\widehat{g}_{kj}$ by using information carried by the correlated entries $\widehat{g}_{qj}$ $(q \neq k)$ in the gain vector $\widehat{\mathbf{g}}_j$, for which the estimated spectral components $\widehat{S}_{qj}$ are unlikely to be all affected by the random noise. The same idea generally holds for other types of randomly localized residual noise. Note that beside the removal of musical noise in silence-dominant segments, the same approach also contributes to restore distorted speech in speech-dominant segments.

Based on these two observations, we aim to reduce the musical or residual noise within $\widehat{S}_{kj}$ by using a calibrated gain function or filter, denoted as $\bar{g}_{kj}$ and computed using the entries of the preliminary gain vector $\widehat{\mathbf{g}}_j$ estimated using (4)–(5). In that regard, our proposed approach is complementary to the temporal smoothing described in (5), as it aims to exploit the gain correlation present along the frequency dimension. For instance, a linear calibration model can be formulated as the following weighted sum,

$$\bar{g}_{kj} = \sum_{q=1}^{K} w_{kq} \widehat{g}_{qj} \tag{6}$$

where w_{kq} is the (k, q)-th entry of a calibration matrix $\mathbf{W} = \left[w_{kq}\right] \in \mathbb{R}^{K \times K}$ and represents the weight given to the q-th entry of the gain vector $\widehat{\mathbf{g}}_j$, in the adjustment of $\widehat{g}_{kj}$. In the case of a linear relationship, the optimal weight values (obtained, e.g., via regularized least-squares) will depend on the correlation between neighboring values of the gain function in the frequency domain. By averaging a large gain value associated with a noise peak at frequency k with smaller gain values in the surrounding of this frequency, we can ideally reduce the energy level of the residual peak in the processed speech below the masking threshold for a human listener. The new enhanced magnitude spectrum is computed as,

$$\widehat{S}_{kj} = \bar{g}_{kj} Y_{kj} \tag{7}$$

where the phase of the noisy speech is unaffected by the enhancement process.

While the calibration approach in (6) leads to improvement in the quality of the enhanced speech, it is restricted to a linear relationship and thus may not fully exploit all the dependencies between nearby values of the gain function g_{kj}. In this work, to overcome this limitation and fully exploit such dependencies, we propose to employ a feed-forward DNN architecture to model both the linear and nonlinear components of the relationship between nearby values of g_{kj} in the frequency domain.

This architecture consists of multiple nonlinear processing layers which together provide a mathematical representation of a highly nonlinear regression function, needed to map a set of preliminary (i.e., less accurate) gain function values at its input, into calibrated (i.e., more robust) gain function values at its output. Each layer, labelled with index $l \in \{1,2, \dots, L\}$, where L is the total number of layers, consists of I_l nodes. The output values of the l-th layer are represented by vector $\mathbf{h}^{(l)} \in \mathbb{R}^{I_l}$ and are expressed as,

$$\mathbf{h}^{(l)} = f_l\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{8}$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{I_l \times I_{l-1}}$ is a linear transformation matrix with (i,j)-th entry $w_{ij}^{(l)}, \mathbf{b}^{(l)} \in \mathbb{R}^{I_l}$ is a bias vector with i-th entry $\mathrm{b}_i^{(l)}$, and $f_l(\bullet)$ represents the activation function of the l-th layer. In this work, the rectified linear unit (ReLU) [6] is used as activation function for the hidden layers ($l = 1, \ldots, L-1$), while the linear function is used for the output layer ($l = L$). In the first layer, $\mathbf{h}^{(0)}$ represents the input vector $\hat{\mathbf{g}}_j \in \mathbb{R}^K$, and in the L-th layer, $\mathbf{h}^{(L)}$ represents the output vector $\mathbf{g}_j \in \mathbb{R}^K$, i.e., the calibrated filter. The complete set of the calibration parameters for the nonlinear DNN model is represented by $\tilde{\mathbf{W}}= \left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)} \,:\, l = 1, \ldots, L\right\}$.

Let $\mathbf{g}_j = \left[g_{1j} g_{2j} \cdots g_{Kj}\right]^{\mathrm{T}}$, where g_{kj} represents the actual Wiener filter gain value at frequency bin k and time frame j, i.e., computed using the clean speech and pure noise data according to (4) and (5), where in the latter equation, the terms $\left(\left[\mathbf{A}_S\mathbf{C}_S\right]_{kj}\right)^2$ and $\left(\left[\mathbf{A}_N\mathbf{C}_N\right]_{kj}\right)^2$ are replaced by $\left|S_{kj}\right|^2$ and $\left|N_{kj}\right|^2$, respectively. During the training stage, the calibration parameters $\mathbf{W}$ are estimated by minimizing the mean-squared error (MSE) between the actual Wiener filter vector $\mathbf{g}_j$, and the calibrated filter $\bar{\mathbf{g}}_j = \left[\bar{g}_{1j}\bar{g}_{2j} \ldots \bar{g}_{Kj}\right]^{\mathrm{T}}$ estimated using the nonlinear combination of the components of the vector $\hat{\mathbf{g}}_j$. Specifically, the cost function is formulated as,

$$E = \frac{1}{N}\sum_{j=1}^{N} \|\bar{\mathbf{g}}_j(\hat{\mathbf{g}}_j, \tilde{\mathbf{W}}) - \mathbf{g}_j\|_2^2 + \lambda \sum_{l=1}^{L} \|\mathbf{W}^{(l)}\|_2^2 \tag{9}$$

where the second term represents the L_2-regularization, $\lambda \geq 0$ is the regularization parameter, and N is the number of frames in the training data.

The calibration parameters $\mathbf{W}$ estimated in the training stage using the DNN will be applied to calibrate the preliminary gain function of the noisy speech in the enhancement stage as shown in Fig. 1. The time-domain enhanced speech signal is finally obtained via inverse STFT followed by the overlap-add method.

3 Multi-filter Model Extension

In the previous section, information about the nonlinear relationship between gain values at different frequency bins was extracted using a single preliminary gain function. In this section, we extend the scope of the calibration process from a single preliminary gain function to multiple gain functions, from which we extract the relationship information. Specifically, by combining a set of gain functions with different properties, we aim to further decrease the variability of the resulting calibrated filter, referred to as a *multi-filter* in the sequel, and consequently reduce the error in the estimated filter. Averaging multiple models, also known as *ensemble method*, is a well-known strategy in machine learning for reducing generalization errors. In particular, when the models are uncorrelated, the expected error can be reduced by

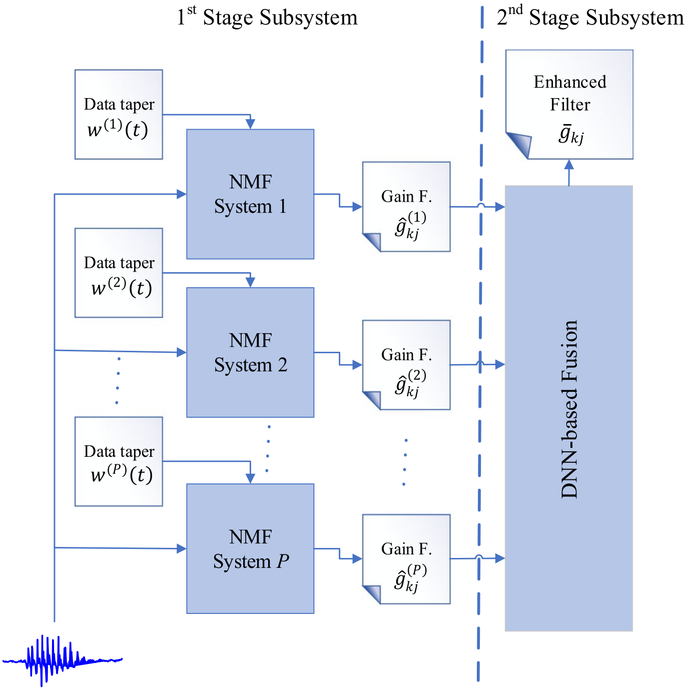

Fig. 3 Block diagram of the speech enhancement system with DNN-based multi-filter

a factor P, where P is the number of combined models [16]. In the present context, several strategies can be followed to design a set of diversified preliminary gain functions. We can, for instance, use a distinct speech feature set as input for each gain function, implement a different speech enhancement algorithm to generate each gain function, or combine both strategies. In this work, we adopt the first strategy wherein multi-tapering is employed to generate the different gain functions using identical NMF systems. Figure 3 depicts the architecture of the proposed speech enhancement system using DNN-based nonlinear fusion to synthesize the desired multi-filter. Below, we describe in more details how the set of preliminary gain functions are designed and combined.

3.1 Designing the Preliminary Gain Function Set

Let $G = \left\{ \hat{g}_{kj}^{(p)} : p = 1 \ldots , P \right\}$ denotes the set of P distinct preliminary gain functions. These gain functions should be designed to provide additional diversity (i.e., richness of information) without incurring a significant computational penalty for the proposed method. In our approach, all the P gain functions $\hat{g}_{kj}^{(p)}$ of Fig. 3 will be generated using the same speech enhancement algorithm, namely the NMF approach, with STFT magnitude data as input, but with the STFT coefficients computed differently for each gain function.

Specifically, for each $p = 1, \dots, P$, the preliminary gain function $\widehat{g}_k^{(p)}$ is estimated as follows,

$$\widehat{g}_{kj}^{(p)} = \frac{\widehat{P}_{kj}^{S}(p)}{\widehat{P}_{kj}^{S}(p) + \widehat{P}_{kj}^{N}(p)} \tag{10}$$

where the power spectra $\widehat{P}_{kj}^{S}(p)$ and $\widehat{P}_{kj}^{N}(p)$ are computed as,

$$\begin{aligned} \widehat{P}_{kj}^{S}(p) &= \tau_{\mathrm{s}} \widehat{P}_{k,j-1}^{S}(p) + \left(1 - \tau_{\mathrm{s}}\right)\left(\left[\mathbf{A}_S^{(p)} \mathbf{C}_S^{(p)}\right]_{kj}\right)^2 \\ \widehat{P}_{kj}^{N}(p) &= \tau_{\mathrm{N}} \widehat{P}_{k,j-1}^{N}(p) + \left(1 - \tau_{\mathrm{N}}\right)\left(\left[\mathbf{A}_N^{(p)} \mathbf{C}_N^{(p)}\right]_{kj}\right)^2 \end{aligned} \tag{11}$$

In (11), $\mathbf{A}_S^{(p)}$ and $\mathbf{A}_N^{(p)}$ are the clean speech and the noise signals dictionaries of the p-th NMF subsystem estimated during the training stage, while $\mathbf{C}_S^{(p)}$ and $\mathbf{C}_N^{(p)}$ are the noisy speech activation matrices estimated during the enhancement stage for the same p-th NMF system. For each p, the dictionaries $\mathbf{A}_S^{(p)}$, $\mathbf{A}_N^{(p)}$ as well as the activation matrices $\mathbf{C}_S^{(p)}$, $\mathbf{C}_N^{(p)}$ of the p-th NMF system in (11) are computed on the basis of so-called p-th tapered STFT coefficients, as explained below.

Let $x_j(t)$ denote the time-domain signal of interest during the j-th frame, where $x \equiv s, n$ or y, respectively, stands for clean speech, pure noise, or noisy speech. We define the p-th tapered STFT coefficient of $x_j(t)$ at frequency bin k as,

$$X_{kj}^{(p)} = \sum_{t=0}^{F-1} w^{(p)}(t) x_j(t) e^{-\jmath 2\pi t k/F} \tag{12}$$

where $w^{(p)}(t)$ is the p-th data taper, $p = 1, \dots, P$. In this work, the data tapers are selected from the sine taper family [39], a set of orthonormal tapers formulated as,

$$w^{(p)}(t) = \sqrt{\frac{2}{F+1}} \sin\left(\frac{\pi p(t+1)}{F+1}\right), \quad t = 0,1, \dots, F-1 \tag{13}$$

where the multiplicative factor $\sqrt{2/(F+1)}$ ensures proper normalization.

We recall that the classical power spectrum estimators use a traditional window function $w(t)$, such as *Hamming* or *Hann* in (12) instead of $w^{(p)}(t)$. By combining a set of diversified gain functions generated using different data tapers, we expect to create a smoother calibrated filter, i.e., a filter with reduced variance. The proposed combination, carried out at the gain function level, can be seen as a late fusion strategy (of combining a set of power spectrum estimates based on different data tapers) compared to the early fusion strategy applied in the multi-tapering method [1, 20] reported in the introduction.

3.2 DNN-Based Fusion Model

Next, we describe the nonlinear model that will be used to fuse the multiple preliminary gain functions into the desired multi-filter. By considering a nonlinear model, we aim to extract richer information between the values of the multiple gain functions at different frequencies, which cannot be captured by a linear model. To verify this conjecture, the results of the fusion based on both linear and nonlinear models will be compared in this study; hence, both models are briefly discussed below.

Let $\hat{\mathbf{g}}_j^{(p)} = \left[\hat{g}_{1j}^{(p)}\hat{g}_{2j}^{(p)} \dots \hat{g}_{Kj}^{(p)}\right]^{\mathrm{T}}$ be the p-th estimated gain vector of the j-th speech frame (using the p-th data taper). We model the enhanced multi-filter $\bar{\mathbf{g}}_j = \left[\bar{g}_{1j}\bar{g}_{2j} \dots \bar{g}_{Kj}\right]^{\mathrm{T}}$, as a functional combination of the P gain vectors $\hat{\mathbf{g}}_j^{(p)}$. For the linear model, the multi-filter value $\bar{g}_{kj}$ can be expressed as,

$$\bar{g}_{kj} = \sum_{p=1}^{P}\sum_{q=1}^{K} w_{kq}^{(p)}\hat{g}_{qj}^{(p)} \tag{14}$$

where $w_{kq}^{(p)}$ is the (k, q)-th entry of a fusion matrix $\mathbf{W}^{(\mathbf{p})}=\left[w_{kq}^{(p)}\right] \in \mathbb{R}^{K\times K}$, associated with the p-th preliminary gain function. As in Section II.B, the collection of fusion matrices $\mathbf{W} = \left\{\mathbf{W}^{(\mathbf{p})} : p = 1, \dots, P\right\}$ can be estimated using a regularized least-squares approach.

For the proposed solution based on the DNN feed forward nonlinear model, the fusion parameter set is equal to $\tilde{\mathbf{W}}= \left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)} : l = 1, \dots, L\right\}$, where $\mathbf{W}^{(l)}$ now represents the weight matrix of the l-th hidden layer of the DNN. These parameters are estimated by minimizing the MSE between the actual Wiener gain vectors $\mathbf{g}_j$ and the estimated gain vectors $\mathbf{g}_j$ obtained by the nonlinear combination of the p gain functions $\hat{\mathbf{g}}_j^{(p)}$. Specifically, the cost function is formulated as,

$$E = \frac{1}{N}\sum_{j=1}^{N} \|\bar{\mathbf{g}}_j(\hat{\mathbf{G}}_j, \tilde{\mathbf{W}}) - \mathbf{g}_j\|_2^2 + \lambda \sum_{l=1}^{L} \|\mathbf{W}^{(l)}\|_2^2 \tag{15}$$

where $\hat{\mathbf{G}}_j = \left[\hat{\mathbf{g}}_j^{(1)^T} \hat{\mathbf{g}}_j^{(2)^T} \dots \hat{\mathbf{g}}_j^{(P)^T}\right]^T \in \mathbb{R}^{PK}$ is the extended gain vector containing the P estimated preliminary gain functions, L is the total number of layers of the DNN, $\lambda \geq 0$ is an L_2-regularization parameter, and N is the number of frames in the training data. As before, the actual gain values, $\mathbf{g}_j$, are computed using the clean speech and pure noise training data.

Note that when $P = 1$, i.e., a single preliminary gain function is used in the fusion stage, the previous equation simplifies to (9), the equation of the calibrated filter.

4 Experimental Results

4.1 Experimental Set-Up

The performance of the proposed systems is evaluated using clean speech from the TSP corpus [23] and noise from the NOISEX dataset [46]. For the clean speech, all adult speakers (11 males and 12 females) are selected. For the noise, two subsets of the NOISEX corpus are selected that will be used for evaluation under matched and unmatched conditions. For the matched condition, *buccaneer 1*, *HF channel*, *babble*, *factory 1*, and *pink* noises are used for both training and testing of the proposed systems. For the unmatched condition, the unseen noises *M109*, *F16* and *Destroyerengine* are only used for testing. The clean speech and noise signal datasets (used in matched condition) are each divided into three subsets: (1) training data, used to estimate the parameters of the models, (2) validation data, used to tune the hyper parameters (such as the number of gain functions in the combination stage), and (3) test data, used for final performance evaluation. The training data consist of approximately 2 min of speech segments for each speaker, as well as 3 min of noise segments. The training data are split in turn into two parts: Train1 and Train2. Train1 is used to train the model of the first stage, i.e., to generate the NMF dictionaries, while Train2 is used to estimate the fusion parameters of the second stage. The validation data consist of 11.5 s of speech for each speaker, and 30 s of noise. The same durations are used for the test partition. The noisy speech is generated by adding the noise to the clean speech to obtain input SNR of 0, 5, and 10 dB. Noisy speech at -5 dB input SNR is also generated for the evaluation in unseen input SNR condition. The audio signals are sampled at 16 kHz.

For the NMF system, we use $M = 80$ basis vectors for the clean speech and for the noise. The temporal smoothing factors are selected as $(\tau_S, \tau_N) = (0.4, 0.9)$. For the STFT analysis, we use a window size of $F = 512$ samples with 75% overlap.

The noise dictionary is estimated using a noise-independent approach, i.e., we estimate a single universal noise dictionary covering all types of noise. During the training of the DNN-based fusion system, Adam [27] is used as the optimizer to minimize the mean square error objective function with a learning rate of 0.0001, and a mini-batch size of 64. Following preliminary experiments where different DNN architectures were evaluated, a $L = 3$ layers network with 256 nodes per layer was selected as it offered the best compromise between performance and complexity. As benchmarks in the performance evaluation of the proposed calibrated filter and multi-filter models for speech enhancement, we use both the conventional NMF-based system, i.e., preliminary gain function (4)–(5) without calibration, and SEGAN [36].

The latter is a speech enhancement system based on a generative adversarial network (GAN); the same network configuration as in [36] is used in this study. During the evaluation, PESQ (perceptual evaluation of speech quality) [22], SDR (signal-to-distortion ratio) [47], SSNR (segmental SNR) and, STOI (short-time objective intelligibility) [42] are used as objective measures for the enhanced

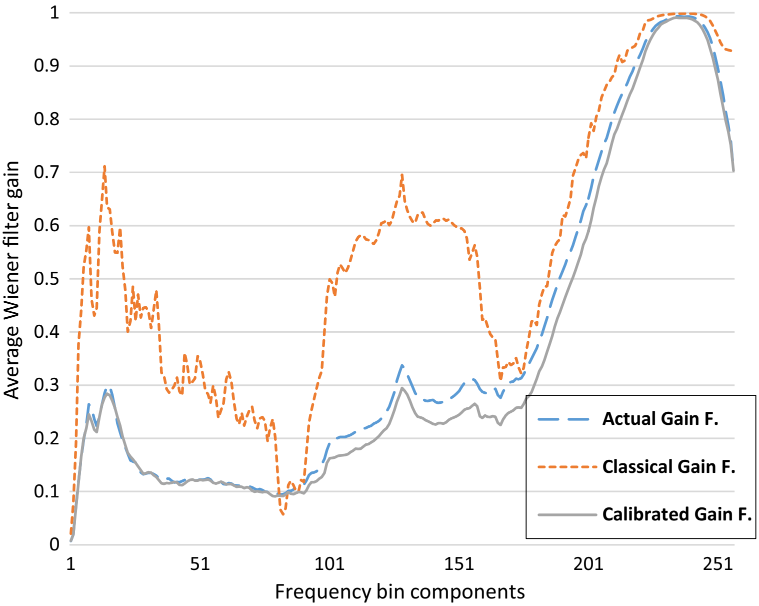

Fig. 4 Average values of the preliminary, calibrated, and the actual gain functions for each frequency bin component, as computed using test data

speech, where a higher value indicates a better speech quality (for PESQ, SDR and SSNR) or speech intelligibility (for STOI). In addition to the above measures of speech quality, we also report the log-kurtosis ratio (LKR) as a measure of the presence of musical noise in the processed speech. The studies in [21, 45] have shown that the human perception of the musical noise is strongly correlated with the (log) kurtosis ratio between the non-speech segments of the noisy and processed speech signal. Specifically, lower LKR values indicate a reduced amount of musical tones in the enhanced speech.

4.2 Calibration and Fusion Effect

The calibration approach developed in Sect. 2.1 aims to correct the noise attenuation level by lowering the estimated preliminary filter value, $\widehat{g}_{kj}$, if there is under estimation of noise, and conversely, by increasing its value if there is an over estimation. To show the effect of the calibration on the preliminary gain function generated by the NMF system, we have plotted in Fig. 4, the average values of the preliminary, calibrated (linear model), and the actual gain functions versus frequency bin, as obtained using test data. The results clearly show the ability of the calibration process to correct the preliminary gain function values, especially in the low to mid frequency interval where the values of the calibrated filter gains are very close to the actual ones. In effect, the implemented NMF in the first stage tends to underestimate the noise power spectrum and leaves

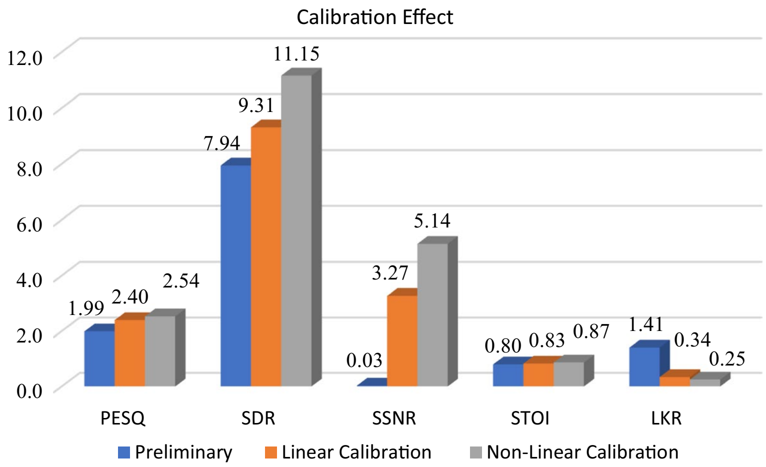

Fig. 5 Calibration effect of the gain function on speech enhancement performance at 5 dB input SNR. The bar charts show average PESQ, SDR, SSNR, STOI, and LKR scores for NMF filter gains (Preliminary), calibrated filter with linear model (Linear Calibration) and calibrated filter with nonlinear model (nonlinear Calibration)

important noise artifacts in the enhanced speech spectrogram, which is subsequently corrected in the second stage of our proposed approach.

Figure 5 shows the calibration effect in terms of PESQ, SDR, SSNR, STOI, and LKR scores for the calibrated filter (with $P = 1$) using both the linear and nonlinear models, and for the conventional (i.e., unprocessed) NMF gain function, where the results are obtained using validation data. As we can observe, the calibration step improves notably the five evaluation scores, reflecting the effect of the adjustment of the gain function shown in Fig. 4. We also observe that the DNN-based nonlinear calibration model performs better than the linear model for the five measures. This suggests that there exists valuable information between gain values extracted by the nonlinear model that are not visible in the plot of Fig. 2.

Next, we illustrate the effect of using multiple gain functions on the speech enhancement performance using validation data. Figure 6 shows that combining more than one individual gain function helps to improve PESQ results independently of the fusion model. The number of gain functions P required to optimize the enhancement results depends on the fusion model. For the linear model, the PESQ results are nearly maximized with a combination of five gain functions, while for the nonlinear model, three are required. We generally find that the nonlinear model gives better results than the linear one, while requiring a smaller number of gain functions. Whereas only the PESQ results are shown in Fig. 6, the same trends are observed when the SDR, SSNR, and STOI are used as evaluation criteria.

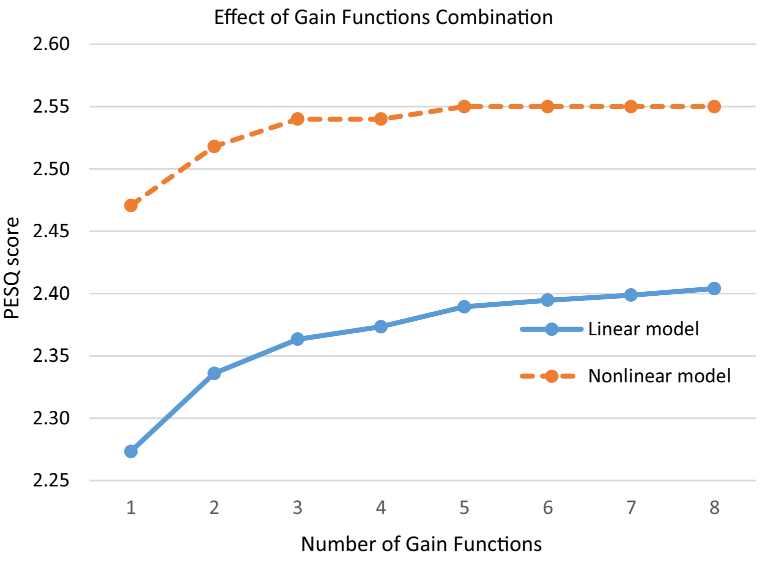

Fig. 6 Impact of increasing the number P of combined gain functions on the PESQ scores at 5 dB input SNR. The plots show the average PESQ scores versus P for the multi-filters derived with linear and nonlinear models

5 Results and Discussion

In this section, the generalization capability of the proposed calibrated filter and the multi-filter models is evaluated using test (unseen) data. Specifically, the calibrated filter (*Calibrated*) and the multi-filter (*Multi-filter*) NMF-based systems are compared with the conventional NMF and the SEGAN systems. For the multi-filter model, the number of data tapers is set to $P = 3$.

Figure 7 illustrates the magnitude spectra of the clean, noisy and enhanced speech for the reference models, i.e., conventional NMF and SEGAN, and the proposed methods with calibrated filter and multi-filter models. In this example, a female speech is degraded with buccaneer noise at the unseen − 5 dB input SNR. We can clearly see that, even though the preliminary gain function, i.e., conventional NMF, removes considerable noise, the low-frequency band still contains important amount of residual noise. While SEGAN removes most of noise at low-frequency band, it leaves the middle-frequency band corrupted by noise. In contrast, the proposed systems with calibrated filter and multi-filter can substantially reduce the remaining low frequency noise left by the preliminary NMF gain without additional signal distortion (corroborating the finding of Fig. 4). In particular, in the case of the multi-filter model the spectrogram is more alike the clean one.

In Figs. 8, 9, 10 and 11, we report the PESQ, SDR, STOI, and LKR results for the matched noise type condition at input SNRs of 0, 5, and 10 dB and for the unseen − 5 dB input SNR. In terms of speech quality, we observe considerable improvements in PESQ, SDR scores with the proposed systems for all noise types and under all input SNR conditions. On average, absolute improvements of about 0.44, and

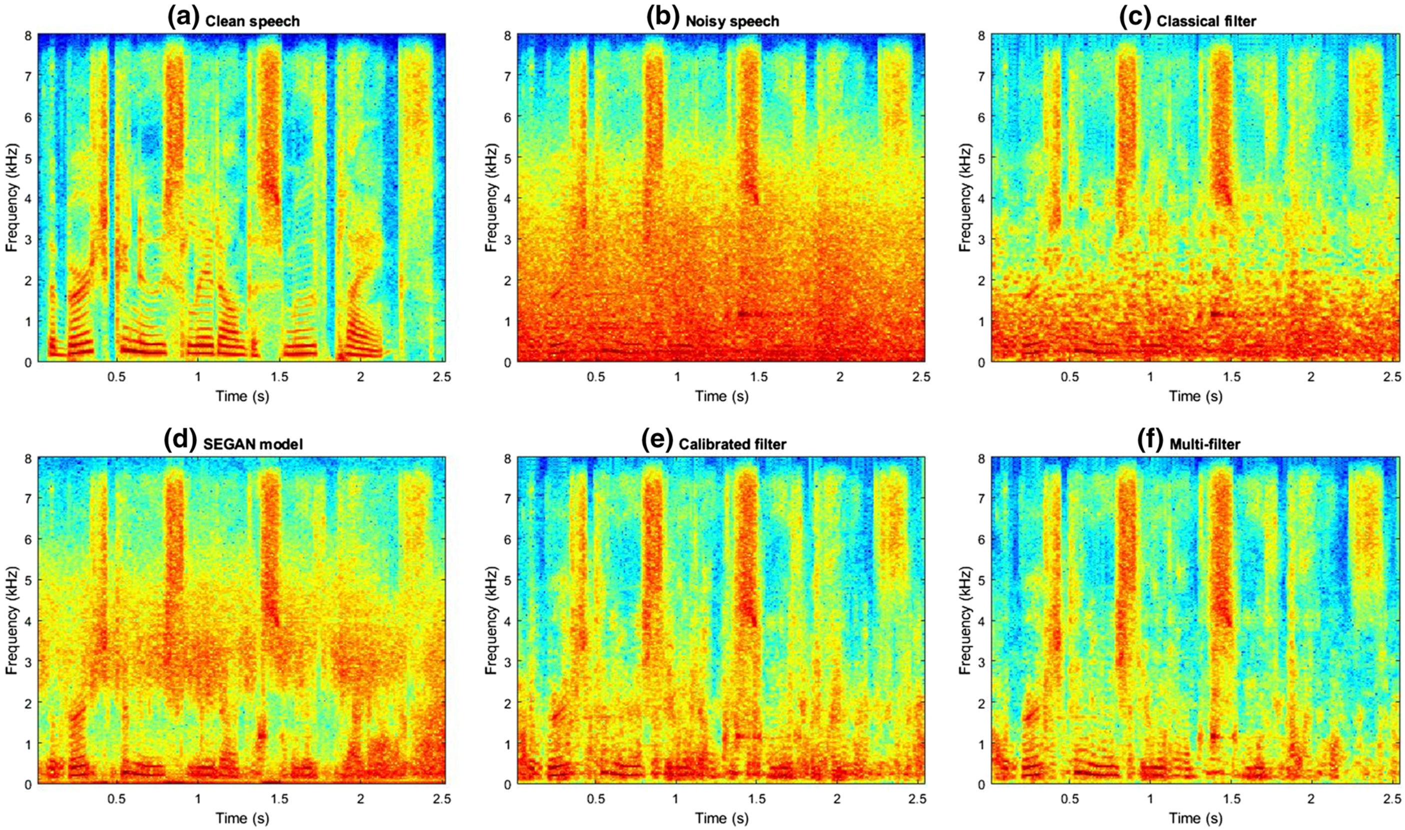

Fig. 7 Spectrogram examples of: **a** female clean speech, **b** noisy speech (clean speech degraded with buccaneer noise at unseen − 5 dB input SNR), and denoised speech using the different types of gain functions, i.e., **c** conventional NMF, **d** SEGAN, **e** calibrated filter, and **f** multi-filter

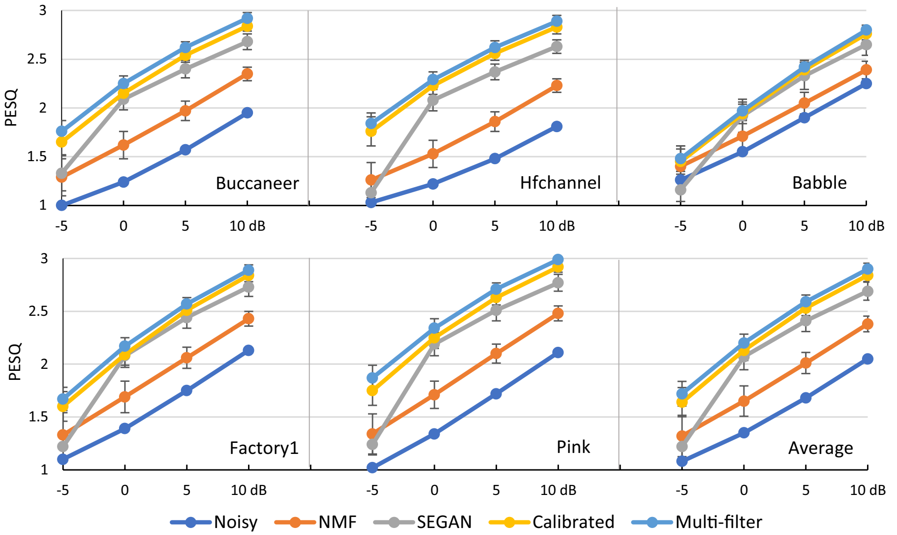

Fig. 8 PESQ results achieved on test data in matched noise condition

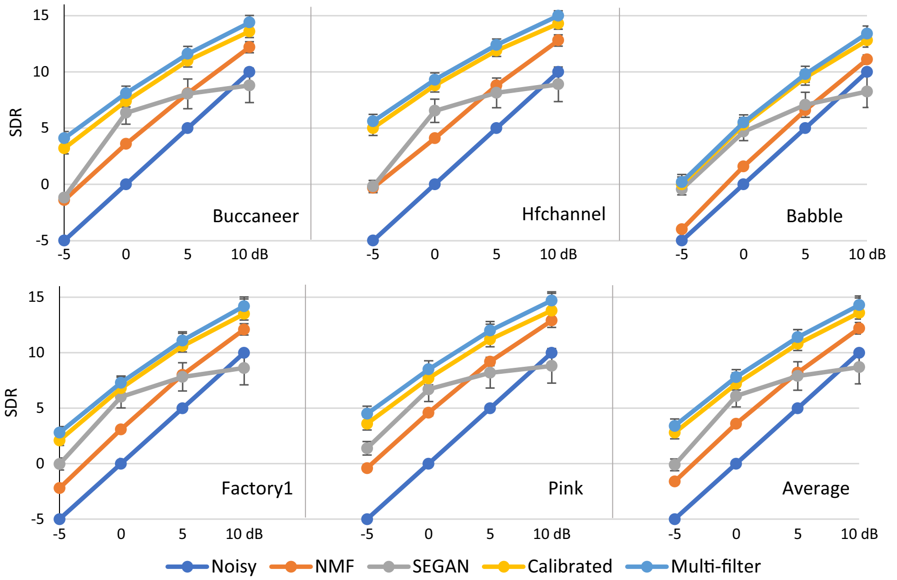

Fig. 9 SDR results achieved on test data in matched noise condition

3.1 dB are achieved for PESQ, and SDR, respectively, for the calibrated filter model, and 0.5, and 3.7 dB for the multi-filter model when compared to conventional NMF system (i.e., with no calibration).

In terms of speech intelligibility, we note that while the conventional NMF system does not improve the STOI score compared to the noisy speech, the calibrated

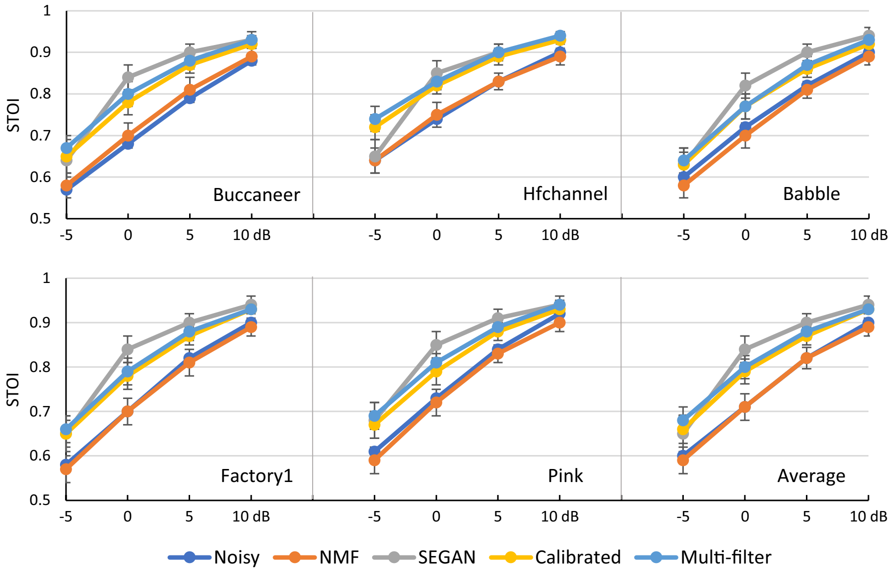

Fig. 10 STOI results achieved on test data in matched noise condition

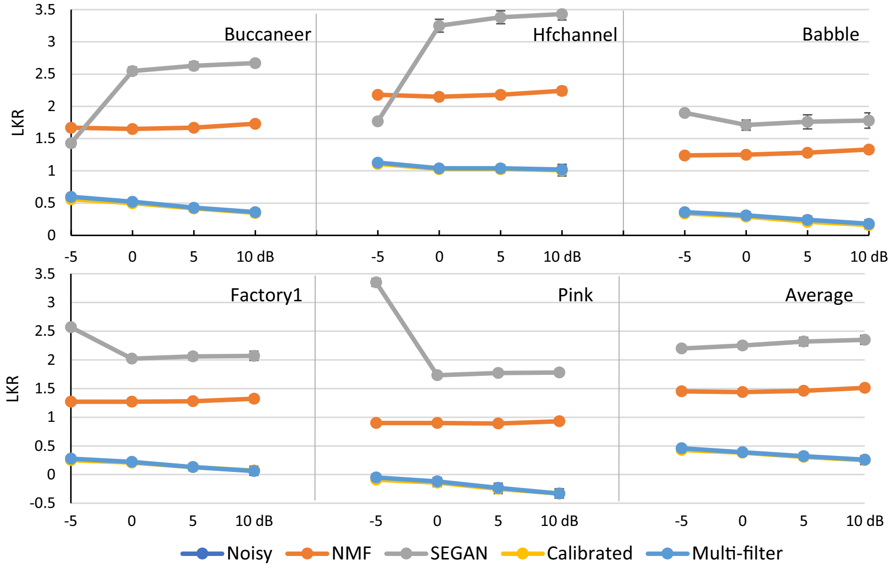

Fig. 11 LKR results achieved on test data in matched noise condition

filter and the multi-filter systems can notably ameliorate the speech intelligibility score with, on average, absolute improvements of about 6% and 7%, respectively.

Interestingly, the gain in speech quality and intelligibility with the proposed calibrated and multi-filter models is achieved without introduction of musical

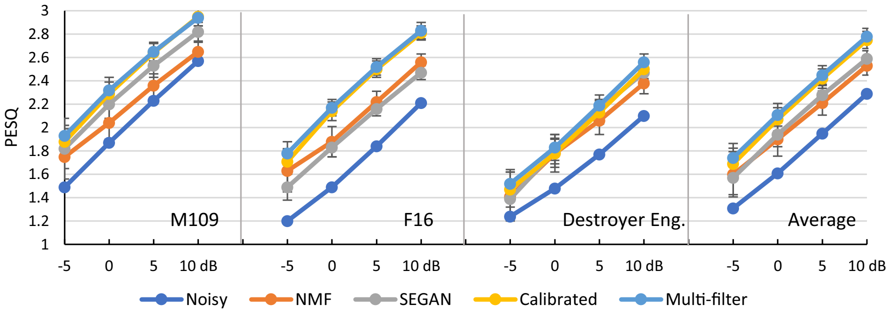

Fig. 12 PESQ results achieved on test data in unmatched noise condition

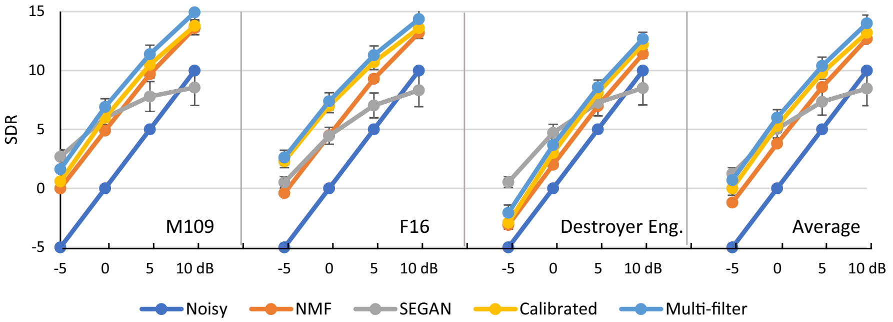

Fig. 13 SDR results achieved on test data in unmatched noise condition

tones artifacts. In effect, the significantly lower LKR results reported for these models in Fig. 11 indicate that the proposed calibration approaches are effective in reducing the amount of musical noise, as initially advocated.

Besides, we also observe that the proposed NMF-based systems with calibrated filter models outperform SEGAN with respect to PESQ, SDR, SSNR, and LKR scores. For the STOI scores, SEGAN is slightly better in matched input SNR conditions for intermediate SNR values (0, and 5 dB).

While the results for the SSNR are not shown in the above figures to improve readability, a similar trend as for SDR has been observed, with an average absolute improvement of about 5.8 dB for the calibrated filter model, and 6.2 dB for the multi-filter model when compared to the conventional NMF system.

Figures 12, 13, 14 and 15 show the PESQ, SDR, STOI, and LKR scores achieved by the various systems at input SNRs of 0, 5, and 10 dB and for the unseen -5 dB input SNR conditions as in the previous tables, but in unmatched (unseen noise) conditions, i.e., with *M109*, *F16*, and *destroyer engine* noises. The proposed calibrated filter models are particular effective in increasing the PESQ (speech quality) and reducing the LKR (musical noise). The results in Figs. 12, 13, 14 and 15 are generally consistent with those made previously under matched noise conditions. We summarize, in Table 1, the average PESQ, SDR, SSNR,

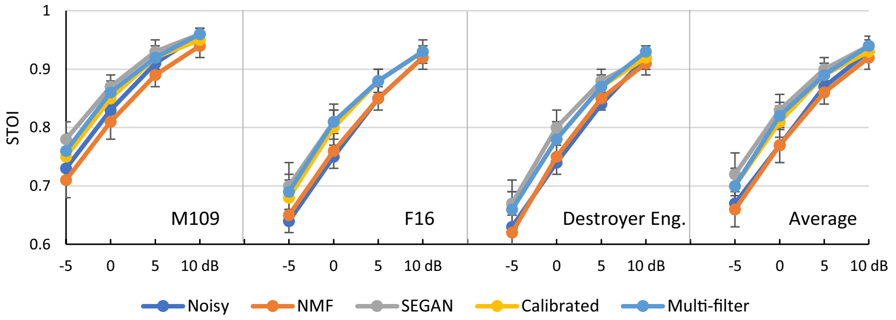

Fig. 14 STOI results achieved on test data in unmatched noise condition

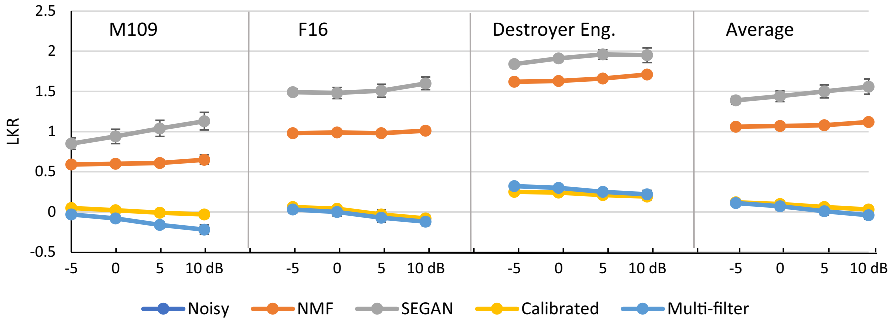

Fig. 15 LKR results achieved on test data in unmatched noise condition

Table 1 Summary of results achieved on test data in matched and unmatched noise conditions

	Noisy	NMF	SEGAN	Calibrated	Multi-filter
Matched conditions					
PESQ	1.54	1.84 ± 0.13	2.10 ± 0.11	2.28 ± 0.09	**2.34 ± 0.08**
SDR	2.5	5.5 ± 0.34	5.6 ± 1.08	8.6 ± 0.57	**9.2 ± 0.65**
SSNR	− 6.8	− 2.7 ± 1.3	0.39 ± 1.0	3.1 ± 1.0	**3.5 ± 0.9**
STOI	0.76	0.75 ± 0.03	0.80 ± 0.03	0.81 ± 0.02	**0.82 ± 0.02**
LKR	–	1.47 ± 0.03	2.31 ± 0.07	0.31 ± 0.06	**0.32 ± 0.05**
Unmatched conditions					
PESQ	1.79	2.06 ± 0.13	2.10 ± 0.10	2.23 ± 0.10	**2.27 ± 0.09**
SDR	2.5	6.0 ± 0.30	5.53 ± 0.97	7.1 ± 0.56	**7.8 ± 0.67**
SSNR	− 6.8	− 2.4 ± 1.4	− 0.98 ± 1.2	0.1 ± 1.2	**0.5 ± 1.2**
STOI	0.81	0.81 ± 0.03	**0.85 ± 0.03**	0.83 ± 0.02	0.84 ± 0.02
LKR	–	1.09 ± 0.03	1.48 ± 0.07	0.08 ± 0.04	**0.04 ± 0.04**

STOI, and LKR scores in matched and unmatched noise conditions, over all noise types and input SNR levels.

Informal subjective listening tests that we have conducted also lead to similar conclusions, indicating that the enhanced speech using the multi-filter model produces the best speech quality and intelligibility, followed by the calibrated-filter and then by SEGAN and the conventional NMF system.

Finally, the proposed multi-filter model has a tractable memory and time complexity. Regarding the first stage subsystem, the NMF algorithm is simple to implement and requires small storage space in comparison with the traditional machine learning methods [7]. For the complexity pertaining to the second stage, we combine the output of a maximum of three NMF systems using a simple fully connected DNN architecture. This DNN subsystem uses an input feature vector of dimension $256\times3=768$ and comprises only two hidden layers, which can be considered as a low-complexity structure compared to deeper DNN structures or more complex architectures such the recurrent neural network (RNN) and convolutional neural network (CNN). Furthermore, for applications where the execution time is the primary concern, we have proposed the calibrated filter model, which provides an interesting compromise between speed and precision. This calibrated model is based on the fine-tuning of a unique NMF system, which provides enhancement performance close to the multi-filter model using three NMF subsystems.

6 Conclusion

We have presented a new two-stage speech enhancement method, specially designed to reduce musical noises without the need for time–frequency localization of the noise peaks. In the first stage of the proposed method, a preliminary gain function is generated using the NMF algorithm. In the second stage, a calibrated gain function, which is more robust to the noise artefacts, is estimated by applying a DNN-based nonlinear mapping function to the preliminary gain function.

To further decrease the variability of the estimated calibrated filter, the DNN-based extraction of frequency dependencies was expanded to a set of preliminary gain functions derived from spectral estimates based on a family of data tapers. The evaluation of the proposed DNN-based calibrated filter models for speech enhancement under different noise types and input SNR has shown substantial improvements in terms of standard speech quality measures compared to the conventional NMF system (with unprocessed gains) and the recently proposed SEGAN system. Finally, the proposed calibrated filter models allow reducing the amount of musical noise in the processed speech without the complex and error-prone task of localizing spurious energy peaks in the spectrogram.

Author Contributions YA and BC conceived and designed the study. YA performed the experiments. YA, BC, and WPZ contributed to the writing, reviewing and editing of the manuscript. All authors read and approved the manuscript.

Funding This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Microsemi Corporation [CRD Grant No. CRDPJ 515072-17].

Availability of Data and Materials Data used in our experiments are all available. The TSP speech database is covered by a permissive Simplified BSD license, and freely available for download at http://www.mmsp.ece.mcgill.ca/Documents/Data/. Noise data can be found at: http://spib.linse.ufsc.br/noise.html.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no competing interests.

References

1. Y. Attabi, H. Chung, B. Champagne, W.-P. Zhu, NMF-based speech enhancement using multitaper spectrum estimation, in *Proceedings of the International Conference on Signals and Systems (ICSigSys)* (2018), pp. 36–41
2. A. Ben Aicha, S. Ben Jebara, Perceptual musical noise reduction using critical bands tonality coefficients and masking thresholds, in *Proceedings of 8th Annual Conference of the International Speech Communication Association* (2007), pp. 822–825
3. S. Ben Jebara, A perceptual approach to reduce musical noise phenomenon with wiener denoising technique, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3 (2006), pp. 49–52
4. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
5. O. Cappé, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. **2**(2), 345–349 (1994)
6. W. Chan, I. Lane, Deep recurrent neural networks for acoustic modelling. arXiv preprint arXiv:1504.01482 (2015)
7. H. Chen, M. Gao, Y. Zhang, W. Liang, X. Zou, Attention-based multi-NMF deep neural network with multimodality data for breast cancer prognosis model. BioMed Res. Int. (2019). https://doi.org/10.1155/2019/9523719
8. H. Chung, E. Plourde, B. Champagne, Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement. Speech Commun. **87**, 18–30 (2017)
9. N. Derakhshan, M. Rahmani, A. Akbari, A. Ayatollahi, An objective measure for the musical noise assessment in noise reduction systems, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2009), pp. 4429–4432
10. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
11. H. Erdogan, J.R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2015), pp. 708–712
12. T. Esch, P. Vary, Efficient musical noise suppression for speech enhancement system, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2009), pp. 4409–4412
13. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
14. T. Gerkmann, R.C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012)
15. Z. Goh, K.-C. Tan, T. Tan, Postprocessing method for suppressing musical noise generated by spectral subtraction. IEEE Trans. Speech Audio Process. **6**(3), 287–292 (1998)
16. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, Cambridge, 2016).
17. H. Gustafsson, S.E. Nordholm, I. Claesson, Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Process. **9**(8), 799–807 (2001)

18. R. Hamon, V. Emiya, L. Rencker, W. Wang, M. Plumbley, Assessment of musical noise using localization of isolated peaks in time-frequency domain, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2017), pp. 696–700
19. Y. Hu, P.C. Loizou, A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. **11**(4), 334–341 (2003)
20. Y. Hu, P.C. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process. **12**(1), 59–67 (2004)
21. T. Inoue, H. Saruwatari, K. Shikano, K. Kondo, Theoretical analysis of musical noise in Wiener filtering family via higher-order statistics, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (2011), pp. 5076–5079
22. ITU-T, *Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ): And Objective Method for End-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Technical Report, 2001
23. P. Kabal, TSP speech database, McGill University, Database Version, vol. 1, no. 0, pp. 09-02, 2002
24. S. Kamath, P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4 (2002), pp. 4164–4164
25. T.G. Kang, K. Kwon, J.W. Shin, and N.S. Kim, NMF-based speech enhancement incorporating deep neural network, in *Proceedings of 15th Annual Conference of the International Speech Communication Association* (2014), pp. 2843–2846
26. M.R. Khan, T. Hasan, M.R. Khan, Iterative noise power subtraction technique for improved speech quality, in *Proceedings of International Conference on Electrical and Computer Engineering* (2008), pp. 391–394
27. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. K. Kwon, J.W. Shin, N.S. Kim, NMF-based speech enhancement using bases update. IEEE Signal Process. Lett. **22**(4), 450–454 (2015)
29. D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing Systems* (2001), pp. 556–562.
30. S. Li, J.-Q. Wang, M. Niu, X.-J. Jing, T. Liu, Iterative spectral subtraction method for millimeter-wave conducted speech enhancement. J. Biomed. Sci. Eng. **3**(2), 187 (2010)
31. J.S. Lim, A.V. Oppenheim, Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)
32. P.C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Cambridge, 2007).
33. R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, K. Kondo, Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Trans. Audio Speech Lang. Process. **20**(7), 2080–2094 (2012)
34. N. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2140–2151 (2013)
35. A. Narayanan, D. Wang, Ideal ratio mask estimation using deep neural networks for robust speech recognition, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), pp. 7092–7096
36. S. Pascual, A. Bonafonte, J. Serrà, SEGAN: speech enhancement generative adversarial network, in *Proceedings of 18th Annual Conference of the International Speech Communication Association* (2017), pp. 3642–3646
37. E. Plourde, B. Champagne, Auditory-based spectral amplitude estimators for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1614–1623 (2008)
38. T.F. Quatieri, R.B. Dunn, Speech enhancement based on auditory spectral change, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1 (2002), pp. 257–260
39. K.S. Riedel, A. Sidorenko, Minimum bias multiple taper spectral estimation. IEEE Trans. Signal Process. **43**(1), 188–195 (1995)
40. P. Scalart, Speech enhancement based on a priori signal to noise estimation, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 2 (1996), pp. 629–632.
41. M.K. Singh, S.Y. Low, S. Nordholm, Z. Zang, Bayesian noise estimation in the modulation domain. Speech Commun. **96**, 81–92 (2018)

42. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
43. D.J. Thomson, Spectrum estimation and harmonic analysis. Proc. IEEE **70**(9), 1055–1096 (1982)
44. R.M. Udrea, N. Vizireanu, S. Ciochina, S. Halunga, Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale. Signal Process. **88**(5), 1299–1303 (2008)
45. Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proceedings of International Workshop on Acoustic Echo and Noise Control* (2008)
46. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
47. E. Vincent, R. Gribonval, C. Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
48. N. Virag, Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. **7**(2), 126–137 (1999)
49. T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1066–1074 (2007)
50. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**, 1702–1726 (2018)
51. D.S. Williamson, Y. Wang, D. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
52. B. Xia, C. Bao, Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. Speech Commun. **60**, 13–29 (2014)
53. Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2015)
54. K. Yamashita, S. Ogata, T. Shimamura, Spectral subtraction iterated with weighting factors, in *Proceedings of IEEE Speech Coding Workshop* (2002), pp. 138–140
55. P.C. Yong, S. Nordholm, H.H. Dam, Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. Speech Commun. **55**(2), 358–376 (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yazid Attabi[1] · **Benoit Champagne**[1] · **Wei-Ping Zhu**[2]

Benoit Champagne
benoit.champagne@mcgill.ca

Wei-Ping Zhu
weiping@ece.concordia.ca

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, QC H3A 0E9, Canada

[2] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada