electronics

MDPI

Article

Supervised Single Channel Speech Enhancement Based on Dual-Tree Complex Wavelet Transforms and Nonnegative Matrix Factorization Using the Joint Learning Process and Subband Smooth Ratio Mask

Md Shohidul Islam, Tarek Hasan Al Mahmud, Wasim Ullah Khan and Zhongfu Ye *

National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei 230026, Anhui, China; shohid7@mail.ustc.edu.cn (M.S.I.); tarek@mail.ustc.edu.cn (T.H.A.M.); kwasim814@mail.ustc.edu.cn (W.U.K.)

* Correspondence: yezf@ustc.edu.cn

Received: 15 February 2019; Accepted: 21 March 2019; Published: 22 March 2019

Abstract: In this paper, we propose a novel speech enhancement method based on dual-tree complex wavelet transforms (DTCWT) and nonnegative matrix factorization (NMF) that exploits the subband smooth ratio mask (ssRM) through a joint learning process. The discrete wavelet packet transform (DWPT) suffers the absence of shift invariance, due to downsampling after the filtering process, resulting in a reconstructed signal with significant noise. The redundant stationary wavelet transform (SWT) can solve this shift invariance problem. In this respect, we use efficient DTCWT with a shift invariance property and limited redundancy and calculate the ratio masks (RMs) between the clean training speech and noisy speech (i.e., training noise mixed with clean speech). We also compute RMs between the noise and noisy speech and then learn both RMs with their corresponding clean training clean speech and noise. The auto-regressive moving average (ARMA) filtering process is applied before NMF in previously generated matrices for smooth decomposition. An ssRM is proposed to exploit the advantage of the joint use of the standard ratio mask (sRM) and square root ratio mask (srRM). In short, the DTCWT produces a set of subband signals employing the time-domain signal. Subsequently, the framing scheme is applied to each subband signal to form matrices and calculates the RMs before concatenation with the previously generated matrices. The ARMA filter is implemented in the nonnegative matrix, which is formed by considering the absolute value. Through ssRM, speech components are detected using NMF in each newly formed matrix. Finally, the enhanced speech signal is obtained via the inverse DTCWT (IDTCWT). The performances are evaluated by considering an IEEE corpus, the GRID audio-visual corpus, and different types of noises. The proposed approach significantly improves objective speech quality and intelligibility and outperforms the conventional STFT-NMF, DWPT-NMF, and DNN-IRM methods.

Keywords: Dual-tree complex wavelet transform (DTCWT); discrete wavelet packet transform (DWPT); stationary wavelet transform (SWT); speech enhancement (SE)

1. Introduction

Speech quality and intelligibility are degraded due to the presence of different types of noise. It is also worsened when environmental and background noises are present. A speech enhancement (SE) approach is developed to reduce the effect of noise, boosting the perceived quality and intelligibility of speech. In the literature, different SE approaches have been formulated based on speech and noise characteristics. Some of them are spectral subtraction [1,2], maximum likelihood spectral amplitude estimation [3], maximum a posteriori spectral amplitude estimation [4], statistical model-based estimation [5], sparseness and temporal gradient regularization method [6], Wiener filtering [7], subband forward algorithm [8], and subspace method [9]. These methods consist of two parts: Noise tracking and signal gain estimation. The noise tracking part tracks the background noise power, whereas the signal gain estimation part estimates the clean speech. Recently, there have been many advances in SE, such as deep neural networks [10,11], deep denoising auto-encoder [12], sparse coding [13–17], and non-negative matrix factorization [18,19]. Most of these methods contain training and testing phases for SE. The training phase prepares a model to characterize the correlation between clean speech and noise; this model is then adopted in the testing phase to recover the clean signal.

Most of the existing methods are based on the short time fourier transforms (STFT). These methods show two problems, including the time-frequency resolution problem, and estimation of the speech signal from the enhanced magnitude spectrum and the noisy phase [20]. Nowadays, some of the research is focused on wavelet-based transforms, such as continuous wavelet transforms (CWT), discrete wavelet transforms (DWT) [21,22], discrete wavelet packet transforms (DWPT) [23–26], and stationary wavelet transforms (SWT) [27]. The CWT provides enormous redundant information and requires a significant amount of computation time and resources. The DWT does not give a better estimation of the critical subband decomposition. The DWPT provides sufficient information both for analysis and synthesis of the original signal, with a significant reduction in the computation time. Moreover, this transform suffers from a lack of shift invariance. Specifically, small shifts in the input signal can cause significant variations in the distribution of energy between coefficients at different levels, resulting in signal reconstruction errors [27]. Both the DWT and DWPT coefficients in each level will be downsampled after filtering. The SWT eliminates the downsampling process at each level to obtain the shift invariance property; furthermore, it is a very redundant transform [27]. In this paper, we focus on the dual-tree complex wavelet transform (DTCWT) based method for speech enhancement. The reason is that it has a shift-invariant property with limited redundancy.

Recently, the authors in [28] proposed an efficient method based on the ratio mask (RM) (continuously valued mask) to overcome outliers (i.e., outlier means the estimate speech magnitude is larger than the mixture's), and adverse magnitude problems. Here, the RM for speech/noise refers to the ratio between the speech/noise magnitude and their combinations in the time-frequency domain. The authors in [28,29] showed that the RM significantly improves speech quality and intelligibility. The objective of the paper is to concatenate a clean subband signal with its RM for decomposing through the NMF. To the best of our knowledge, this has not been studied yet. It is noted that we measure the ideal RMs (IRMs) for clean speech and noise from their mixtures in the training phase, whereas the enhanced RMs (ERMs) of speech are measured in the enhancement phase. NMF was directly applied to the matrix of the subband signal in the wavelet domain, similar to [24]. As a result, some errors occur during decompositions of the signal using NMF.

Furthermore, we use the auto-regressive moving average (ARMA) filtering process in the wavelet domain of smooth decomposition. Different types of RM have been studied in the literature to increase the quality of the estimated signal, such as an ideal or standard ratio mask (iRM or sRM), square root ratio mask (srRM), and binary ratio mask (bRM). The bRM provides a hard decision, i.e., its mask value is 0 if it is dominated by noise, whereas a mask value is 1 when it is dominated by speech. As a result, it removes the background noise imperfectly from the subband signal dominated by the speech, severely degrading the hearing quality. Contrarily, the sRM in [29] considered the soft decision where the mask values continuously vary from 0 to 1 instead of the hard decision by the bRM. The srRM has shown a better enhancing performance over the bRM in [30] because srRM is more suitable than sRM to keep both the speech and noise energy, and provides a higher mask value in the range of 0 to 1 than sRM. The higher mask value has a comparatively weakened ability to mask the noise; besides, it has an excellent ability to keep the speech components. To this place, we apply the subband smooth ratio mask (ssRM) to take advantage of the combined use of the sRM and srRM. The key contributions of this work can be summarized as follows:

Firstly, we propose a novel speech enhancement approach that adheres to the DTCWT and NMF with the Kullback-Leibler (KL) divergence cost function. Secondly, we derive the new matrix by applying the framing scheme, where the row corresponds to the frame number and the column to the frame length. Thirdly, we calculate the RMs for speech and noise from their mixtures and concatenate them with the previously formed matrix of speech and noise, respectively. Then, we take the absolute value and employ the ARMA filter to obtain smooth decomposition instead of direct use of the NMF. Finally, we propose the new ssRM by combining the sRM and srRM to enhance the performance of the speech signal.

The rest of this paper is organized as follows. Section 2 provides the formulated model of the proposed method. In Section 3, a short overview of NMF with cost functions is given. In Section 4, a brief explanation about DTCWT is presented. Section 5 provides a brief description of the DTCWT-NMF based speech enhancement method. Section 6 presents speech and noise databases with experimental conditions and discussion. Section 7 concludes the paper followed by the references.

2. Problem Formulation

Let us consider the observed speech signal, $\mathrm{x(t)}$, as

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

where $\mathrm{s(t)}$ is a clean speech signal and $\mathrm{n(t)}$ is a noise signal. Then, the DTCWT of the noisy signal in Equation (1) can be written as:

$$\mathrm{DTCWT\{x(t)\} = DTCWT\{s(t)\} + DTCWT\{n(t)\}} \tag{2}$$

$$\mathrm{DTcwt_x = DTcwt_s + DTcwt_n} \tag{3}$$

where DTCWT represents the dual-tree complex wavelet transforms, $\mathrm{DTcwt_x}$ signifies the RMs of the wavelet coefficients of the observed signal, $\mathrm{DTcwt_s}$ denotes the RMs of the wavelet coefficients of the clean speech signal, and $\mathrm{DTcwt_n}$ indicates the RMs of the wavelet coefficients of the noise. Here, $\mathrm{DTcwt_s}$ and $\mathrm{DTcwt_n}$ are unknown RMs of the wavelet coefficients and need to be estimated using the RMs of the noisy wavelet coefficients and the RMs of the clean training speech and the noise wavelet coefficients via NMF and the subband smooth ratio mask. Then, the enhanced time domain speech signal is obtained as:

$$\tilde{\mathrm{s}}\mathrm{(t) = DTCWT^{-1}}(\widetilde{DT}cwt_S) \tag{4}$$

where $\tilde{\mathrm{s}}\mathrm{(t)}$ is the estimated time domain clean speech signal, $\mathrm{DTCWT^{-1}}$ represents the IDTCWT, and $\widetilde{DT}cwt_S$ is the estimated clean speech RMs of the wavelet coefficients.

3. Non-Negative Matrix Factorization (NMF)

NMF is an approach that is used for decomposing any nonnegative matrix, X, into a non-negative basis matrix, W, and a nonnegative weight matrix, H. Thus, we have:

$$\mathrm{X = WH} \tag{5}$$

where each column vector of matrix X is estimated by a weighted linear combination of the basis vectors, which are the columns of W and the weights for the basis vectors that appear in the corresponding columns of H. Nonnegative basis vectors of matrix W are optimized to allow the data matrix, X, to be estimated as a positive linear combination of its constituent vectors. The matrices, W and H, are approximated by using the following optimization problem as:

$$\min \mathrm{C(X||WH)} \tag{6}$$

where both matrices, W and H, are nonnegative. C represents the cost function, which approximates the distance between X and WH. Two types of cost function are reported in [31]. The first one is the Euclidean distance (ED) cost function, given as follows:

$$\mathrm{C_{ED}} = \min(||\mathrm{X} - \mathrm{WH}||_2^2) \tag{7}$$

$$= \sum_{\mathrm{i,j}} (\mathrm{X_{i,j}} - (\mathrm{WH})_{\mathrm{i,j}})^2 \tag{8}$$

The matrices, W and H, can be computed through the following iterative updates as:

$$\mathrm{W} \leftarrow \mathrm{W}\frac{\mathrm{XH^T}}{\mathrm{WHH^T}} \tag{9}$$

$$\mathrm{H} \leftarrow \mathrm{H}\frac{\mathrm{W^TX}}{\mathrm{W^TWH}} \tag{10}$$

The second one is the KL divergence cost function, given as follows:

$$\mathrm{C_{KL}} = \min \mathrm{D}(\mathrm{X}||\mathrm{WH}) \tag{11}$$

$$= \sum_{\mathrm{i,j}} (\mathrm{X_{i,j}} \log \frac{\mathrm{X_{i,j}}}{(\mathrm{WH})_{\mathrm{i,j}}} - \mathrm{X_{i,j}} + (\mathrm{WH})_{\mathrm{i,j}}) \tag{12}$$

The KL cost function works well in audio source separation, and therefore, we will consider it in our new approach. The matrices, W and H, are obtained through the following iterative updates:

$$\mathrm{W} \leftarrow \mathrm{W} \otimes \frac{\frac{\mathrm{X}}{\mathrm{WH}}\mathrm{H^T}}{1\mathrm{H^T}} \tag{13}$$

$$\mathrm{H} \leftarrow \mathrm{H} \otimes \frac{\mathrm{W^T}\frac{\mathrm{X}}{\mathrm{WH}}}{\mathrm{W^T}1} \tag{14}$$

where 1 represents a matrix of ones with the same size as X, and the division and the operator, ⊗, are element-wise division and multiplication, respectively.

4. Dual-Tree Complex Wavelet Transform (DTCWT)

The DWT has been well adopted in recent decades in many signal processing applications. This is due to its ability to provide an efficient time-frequency analysis of signals. However, despite these efficiencies, the DWT suffers from the following drawbacks:

- Shift variance.
- Oscillation.
- Aliasing.
- Lack of orientation.

The SWT solves the lack of shift invariance problem by eliminating downsampling operators, but it is a very highly redundant transform. It represents a redundancy as much as N × J (N is the length of the input signal and J is the maximum number of decomposition levels) and, as a result, denoising procedures consume considerable computation time. One solution for the shift-invariant wavelet transform without high redundancy is the complex wavelet transform, which takes advantage of complex-valued filters instead of real ones. For perfect reconstruction properties, the complex filters must be analytic. However, the successful design of an analytic complex-valued filter for implementation in filter banks is difficult. Kingsbury introduced a more computationally efficient approach for a shift invariance transform; the DTCWT has the following properties [32]:

- Approximate shift invariance.
- Perfect reconstruction.
- Limited redundancy, independent of the number of levels, it has 2:1 redundancy.

- Efficient order—N computation, only twice the DWT.

The two real wavelet transforms use two different sets of filters, with each satisfying the PR conditions. The two sets of filters are jointly designed so that the overall transform is approximately analytic [33]. According to its name, the DTCWT uses two wavelet trees. The input signal is applied to the two trees, and it is decomposed into four subbands. $r_0(n)$, $r_1(n)$ (real part) and $i_0(n)$, $i_1(n)$ (imaginary part) are the low pass and high pass filters of the two wavelet trees, respectively. The two-level decomposition of the DTCWT is shown in Figure 1a, and the filter bank for implementing the inverse of DTCWT (IDTCWT) is shown in Figure 1b. For the implementation of the DTCWT, no complex arithmetic is needed because the filters themselves are real. The number of coefficients in the DTCWT at each level is two times DWT so that the DTCWT is a redundancy, and compared with SWT it is small. So, to analyze and synthesize speech signals, the DTCWT is a powerful mathematical tool.

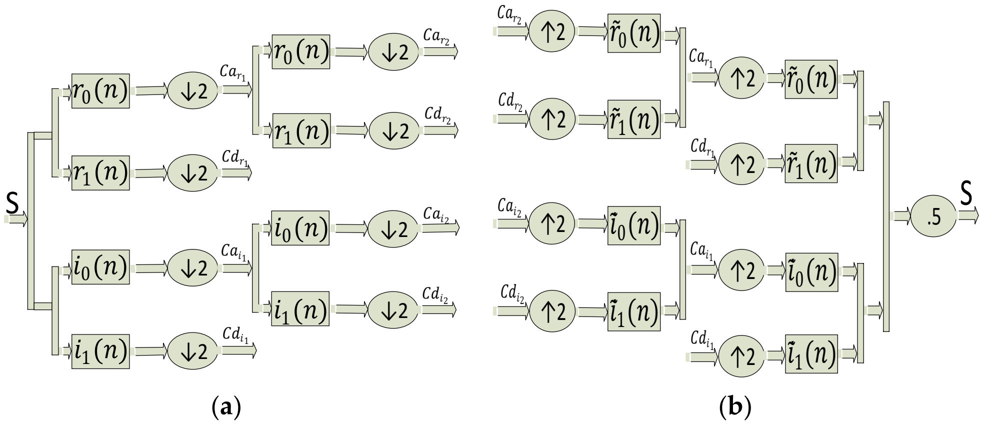


Figure 1. The two-level filter bank implementation of (**a**) DTCWT and (**b**) IDTCWT.

5. DT-CWT-NMF Based Speech Enhancement System

The STFT-NMF and DWPT-NMF based speech enhancement systems were studied in [24,34]. The DNN that is used to estimate the ideal ratio mask (IRM) for speech separation (DNN-IRM) was studied in [29]. The DNN has three hidden layers, each having 1024 rectified linear (ReLU) hidden units. The input layer is given the following set of complementary features that are extracted from a 64-channel Gammatone filterbank: Amplitude modulation spectrogram (AMS), relative spectral transform, and perceptual linear prediction (RASTA-PLP); mel-frequency cepstral coefficients (MFCC); and the cochleagram response, as well as their deltas. The standard backpropagation algorithm coupled with dropout regularization (dropout rate 0.2) is used to train the networks and the adaptive gradient descent is also used along with a momentum term as the optimization technique. A momentum rate of 0.5 is used for the first five epochs, after which the rate increases to and is kept at 0.9. The DNNs are trained to predict the desired outputs across all frequency bands, and the mean squared error (MSE) is used as the cost (loss) function. In this paper, we propose a novel speech enhancement system that adopts DTCWT and NMF with ssRM. Figure 2 depicts the main structure of our proposed speech enhancement system, and it consists of two phases: The training phase and test phase.

5.1. Training Phase

In the training phase, the clean speech signal, $s(t)$, noise, $n(t)$, and noisy speech, $sn(t)$, pass through the DTCWT and yield a set of subband signals, i.e., $\{s^J_{b,t}\}$, $\{n^J_{b,t}\}$, and $\{sn^J_{b,t}\}$, where J represents the level of DTCWT, b denotes the subband index, and t indicates the tree level. The standard deviation of each subband signal is calculated and preserved. The overlapping framing scheme is implemented in each subband signal to create the subband signal matrix, where the column corresponds to the frame number and the row corresponds to the frame length, i.e., $S_{b,t}^{J^{Train}}$, $N_{b,t}^{J^{Train}}$, and $SN_{b,t}^{J^{Train}}$.

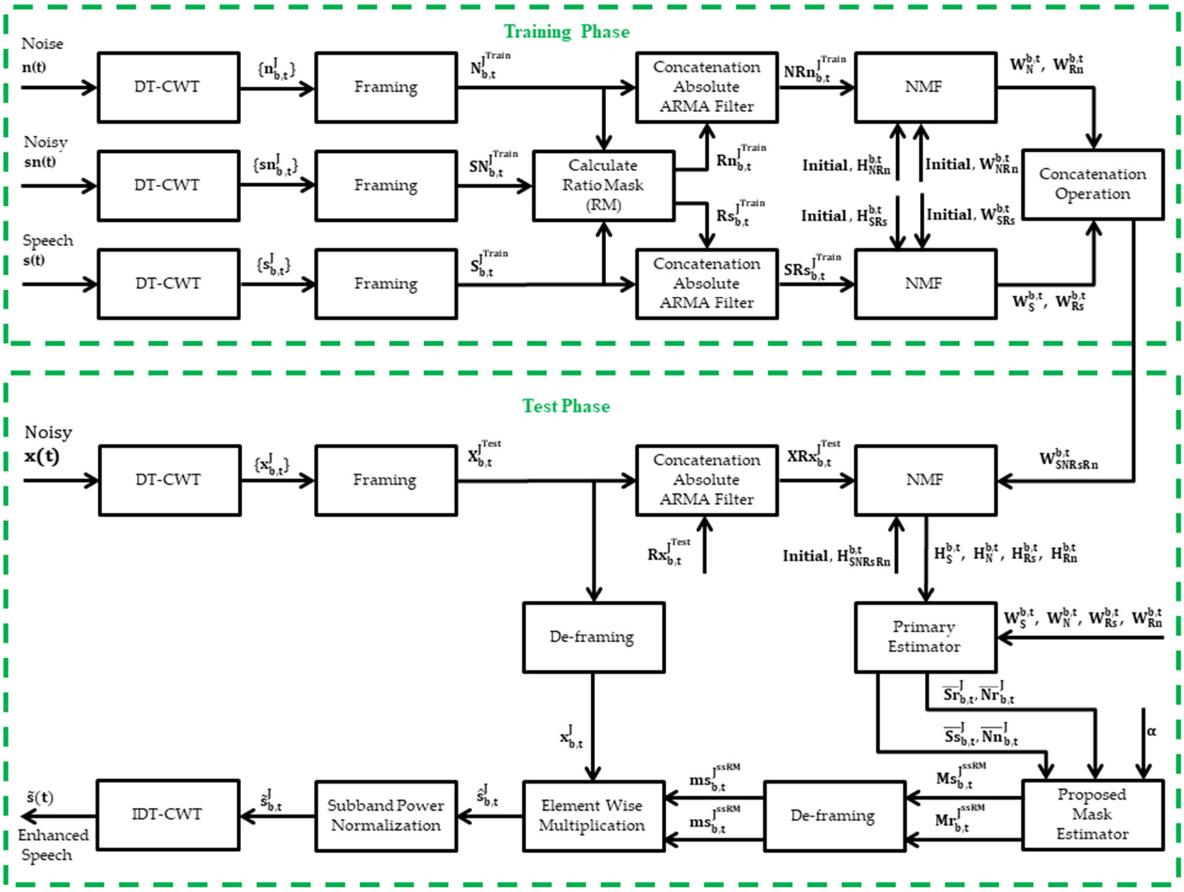


Figure 2. Block diagram of the proposed speech enhancement system.

We calculate the RMs for speech, $Rs_{b,t}^{J^{Train}}$, and noise, $Rn_{b,t}^{J^{Train}}$, from their mixture, $SN_{b,t}^{J^{Train}}$, by using Equations (15) and (16). Then, the RMs are concatenated with their corresponding speech, $S_{b,t}^{J^{Train}}$, and noise, $N_{b,t}^{J^{Train}}$, by using Equations (17) and (18):

$$Rs_{b,t}^{J^{Train}} = S_{b,t}^{J^{Train}} ./SN_{b,t}^{J^{Train}} \tag{15}$$

$$Rn_{b,t}^{J^{Train}} = N_{b,t}^{J^{Train}} ./ SN_{b,t}^{J^{Train}} \tag{16}$$

$$SRs_{b,t}^{J^{Train}} = \begin{bmatrix} S_{b,t}^{J^{Train}} \\ Rs_{b,t}^{J^{Train}} \end{bmatrix} \tag{17}$$

$$NRn_{b,t}^{J^{Train}} = \begin{bmatrix} N_{b,t}^{J^{Train}} \\ Rn_{b,t}^{J^{Train}} \end{bmatrix} \tag{18}$$

We consider only the absolute value of these concatenated matrices to removing negativity, i.e., $|SRs_{b,t}^{J^{Train}}|$ and $|NRn_{b,t}^{J^{Train}}|$. Instead of using NMF directly, to decompose the nonnegative matrices, $|SRs_{b,t}^{J^{Train}}|$ and $|NRn_{b,t}^{J^{Train}}|$, we generate the new matrices by applying the second-order ARMA filtering scheme reported in [35] to the nonnegative matrices for smooth decomposition. We then analyzed the matrix, $SRs_{b,t}^{J^{Train}}$, via NMF, such that $SRs_{b,t}^{J^{Train}} \approx W_{SRs}^{b,t} H_{SRs}^{b,t}$, where the initial values of $W_{SRs}^{b,t}$ and $H_{SRs}^{b,t}$ are initialized by positive random numbers. The clean speech basis matrix, $W_{SRs}^{b,t}$ (i.e., $W_{S}^{b,t}$ for the subband speech signal and $W_{Rs}^{b,t}$ for the subband speech ratio mask are learned jointly via NMF) can be approximated by minimizing the distance between $\mathrm{SRs}_{b,t}^{J^{\mathrm{Train}}}$ and $W_{\mathrm{SRs}}^{b,t}H_{\mathrm{SRs}}^{b,t}$ using Equation (12) with the help of Equations (13) and (14). Similarly, the noise basis matrix, $W_{\mathrm{NRn}}^{b,t}$ (i.e., $W_{N}^{b,t}$ for subband noise and $W_{\mathrm{Rn}}^{b,t}$ for the subband noise ratio mask are learned jointly via NMF), associated with the subband b is derived by using the noise in the training data set. Finally, these matrices are concatenated and obtain the final concatenated matrix as follows:

$$W_{\mathrm{SNRsRn}}^{b,t} = [W_{S}^{b,t}\ W_{N}^{b,t}\ \ W_{\mathrm{Rs}}^{b,t}\ \ W_{\mathrm{Rn}}^{b,t}] \tag{19}$$

The proposed algorithm for the training phase is as follows:

a) Calculate the wavelet coefficients, $\{s_{b,t}^{J}\}$ and $\{n_{b,t}^{J}\}$, from the speech and noise training data via DTCWT, as well as the wavelet coefficients, $\{sn_{b,t}^{J}\}$, of the train noisy signal.

b) Apply the framing scheme on the wavelet coefficients and obtain the matrices, $S_{b,t}^{J}$, $N_{b,t}^{J}$, and $SN_{b,t}^{J}$, for training speech, noise, and noisy signal, respectively.

c) Calculate $\mathrm{Rs}_{b,t}^{J^{\mathrm{Train}}}$ and $\mathrm{Rn}_{b,t}^{J^{\mathrm{Train}}}$ according to Equations (15) and (16) and arrange the speech, $\mathrm{SRs}_{b,t}^{J^{\mathrm{Train}}}$, and noise, $\mathrm{NRn}_{b,t}^{J^{\mathrm{Train}}}$, training sets according to Equations (17) and (18).

d) Take the absolute value and apply the ARMA filter to the $\mathrm{SRs}_{b,t}^{J^{\mathrm{Train}}}$ and $\mathrm{NRn}_{b,t}^{J^{\mathrm{Train}}}$.

e) Decompose the matrices, $I_{b,t}^{J^{\mathrm{Train}}} \approx W_{I}^{b,t}H_{I}^{b,t}$, for $I = \mathrm{SRs}, \mathrm{NRn}$ using NMF with the KL cost function and prepare the matrix, $W_{\mathrm{SNRsRn}}^{b,t}$, using Equation (19).

5.2. Test Phase

In the test phase, the noisy speech signal, $x(t)$, passes through the DTCWT and provides a set of noisy subband signals, i.e., $\{x_{b,t}^{J}\}$. Applying the process above, we obtain the nonnegative matrix as:

$$\mathrm{XRx}_{b,t}^{J^{\mathrm{Test}}} = \left[\, X_{b,t}^{J^{\mathrm{Test}}} \quad \mathrm{Rx}_{b,t}^{J^{\mathrm{Test}}} \right] \tag{20}$$

where $\mathrm{Rx}_{b,t}^{J^{\mathrm{Test}}}$ is the one matrix defined test ratio mask and is the same size as the matrix, $X_{b,t}^{J^{\mathrm{Test}}}$. The NMF is used again to decompose the nonnegative matrix, $\mathrm{XRx}_{b,t}^{J^{\mathrm{Test}}}$, such that $\mathrm{XRx}_{b,t}^{J^{\mathrm{Test}}} \approx W_{\mathrm{SNRsRn}}^{b,t}H_{\mathrm{SNRsRn}}^{b,t} = W_{S}^{b,t}H_{S}^{b,t} + W_{N}^{b,t}H_{N}^{b,t} + W_{\mathrm{Rs}}^{b,t}H_{\mathrm{Rs}}^{b,t} + W_{\mathrm{Rn}}^{b,t}H_{\mathrm{Rn}}^{b,t}$, where $H_{\mathrm{SNRsRn}}^{b,t}$ represents the activation matrix, and $H_{S}^{b,t}$, $H_{N}^{b,t}$, $H_{\mathrm{Rs}}^{b,t}$, and $H_{\mathrm{Rn}}^{b,t}$ denote the speech, noise, speech ratio mask, and noise ratio mask portions of $H_{\mathrm{SNRsRn}}^{b,t}$, respectively. The activation matrix, $H_{\mathrm{SNRsRn}}^{b,t}$ (i.e., $H_{S}^{b,t}$, $H_{N}^{b,t}$, $H_{\mathrm{Rs}}^{b,t}$, and $H_{\mathrm{Rn}}^{b,t}$ matrices are learned jointly via NMF), can be approximated by applying Equations (12) and (14), where the initial value of $H_{\mathrm{SNRsRn}}^{b,t}$ is initialized by a positive random number and the value of $W_{\mathrm{SNRsRn}}^{b,t}$ is fixed, as generated during the training phase. We obtain the rough estimate of the clean speech signal, $(\bar{\mathrm{S}}\mathrm{s}_{b,t}^{J})$, in the noisy test signal, $(X_{b,t}^{J})$, by multiplying the enhanced subband speech basis matrix, $(W_{S}^{b,t})$, with its corresponding activation matrix, $(H_{S}^{b,t})$, and the rough estimate of the clean speech signal, RM $(\bar{\mathrm{S}}\mathrm{r}_{b,t}^{J})$, in the noisy test signal, $(X_{b,t}^{J})$, by multiplying the enhanced RM subband basis matrix, $(W_{\mathrm{Rs}}^{b,t})$, with its corresponding activation matrix, $(H_{\mathrm{Rs}}^{b,t})$, as:

$$\bar{\mathrm{S}}\mathrm{s}_{b,t}^{J} = W_{S}^{b,t}H_{S}^{b,t} \tag{21}$$

$$\bar{\mathrm{S}}\mathrm{r}_{b,t}^{J} = W_{\mathrm{Rs}}^{b,t}H_{\mathrm{Rs}}^{b,t} \tag{22}$$

Similarly, we get rough estimates of the noise, $(\overline{\mathrm{Nn}}_{b,t}^{J})$, and noise ratio mask, $(\overline{\mathrm{Nr}}_{b,t}^{J})$, in the noisy test signal, $(X_{b,t}^{J})$, as:

$$\overline{\mathrm{Nn}}_{b,t}^{J} = W_{N}^{b,t}H_{N}^{b,t} \tag{23}$$

$$\overline{\mathrm{Nr}}_{b,t}^{J} = W_{\mathrm{Rn}}^{b,t}H_{\mathrm{Rn}}^{b,t} \tag{24}$$

The algorithm for the test phase is as follows:

a) Extract the wavelet coefficients, $\{x_{b,t}^{J}\}$, from the noisy test signal via DTCWT.
b) Apply the framing scheme on the wavelet coefficients, $\{x_{b,t}^{J}\}$, and obtain the matrices, $X_{b,t}^{J^{Test}}$.
c) Prepare the noisy test matrix, $XRx_{b,t}^{J^{Test}}$, using the concatenation and ARMA filtering operations on the matrix, $X_{b,t}^{J^{Test}}$, and RM matrix, $Rx_{b,t}^{J^{Test}}$ (i.e., $Rx_{b,t}^{J^{Test}}$ is the one's matrix and is the same size as the matrix, $X_{b,t}^{J^{Test}}$).
d) Learn the weight matrix, $H_{SNRsRn}^{b,t}$, jointly from the noisy test signal, $XRx_{b,t}^{J^{Test}}$, while keeping the basis matrix, $W_{SNRsRn}^{b,t}$, fixed: $XRx_{b,t}^{J^{Test}} \approx W_{SNRsRn}^{b,t}H_{SNRsRn}^{b,t}$.
e) Find the rough estimate of speech, speech RM, noise, and noise RM in the noisy test signal as $\bar{I}_{b,t}^{J} \approx W_{K}^{b,t}H_{K}^{b,t}$ since $I = Ss, Sr, Nn, Nr$ and $K = S, Rs, N, Rn$.

5.3. Subband Smooth Ratio Mask (ssRM)

We can directly use the rough estimate of speech, $\bar{Ss}_{b,t}^{J}$, and noise, $\overline{Nn}_{b,t}^{J}$, or speech RM, $\bar{Sr}_{b,t}^{J}$, and noise RM, $\overline{Nr}_{b,t}^{J}$, but the noisy signal, $X_{b,t}^{J}$, is not exactly equal to the sum of the above two estimates. To make it error free, the mask for speech, $Ms_{b,t}^{J}$, and speech RM, $Msrm_{b,t}^{J}$, are measured as follows:

$$Ms_{b,t}^{J} = (\bar{Ss}_{b,t}^{J})^{p}./((\bar{Ss}_{b,t}^{J})^{p} + (\overline{Nn}_{b,t}^{J})^{p}) \tag{25}$$

$$Msrm_{b,t}^{J} = (\bar{Sr}_{b,t}^{J})^{p}./((\bar{Sr}_{b,t}^{J})^{p} + (\overline{Nr}_{b,t}^{J})^{p}) \tag{26}$$

where p is a parameter, which possesses values greater than zero. The elements of matrices, $Ms_{b,t}^{J}$ and $Msrm_{b,t}^{J}$, vary from zero to one, and different values of p indicate different types of masks. Thus, the sRM $Ms_{b,t}^{J^{sRM}}$ and srRM $Ms_{b,t}^{J^{srRM}}$ for speech are defined in Equations (27) and (28), respectively. Additionally, the sRM $Mr_{b,t}^{J^{sRM}}$ and srRM $Mr_{b,t}^{J^{srRM}}$ for speech RM are defined in Equations (29) and (30), respectively:

$$Ms_{b,t}^{J^{sRM}} = \bar{Ss}_{b,t}^{J}./(\bar{Ss}_{b,t}^{J} + \overline{Nn}_{b,t}^{J}) \tag{27}$$

$$Ms_{b,t}^{J^{srRM}} = \sqrt{\bar{Ss}_{b,t}^{J}}./(\sqrt{\bar{Ss}_{b,t}^{J}} + \sqrt{\overline{Nn}_{b,t}^{J}}) \tag{28}$$

$$Mr_{b,t}^{J^{sRM}} = \bar{Sr}_{b,t}^{J}./(\bar{Sr}_{b,t}^{J} + \overline{Nr}_{b,t}^{J}) \tag{29}$$

$$Mr_{b,t}^{J^{srRM}} = \sqrt{\bar{Sr}_{b,t}^{J}}./(\sqrt{\bar{Sr}_{b,t}^{J}} + \sqrt{\overline{Nr}_{b,t}^{J}}) \tag{30}$$

where "√." and "./" represent the element-wise square root and division operators, respectively. Since the sRM provides a soft decision, where the mask values continuously vary from 0 to 1, contrarily, the bRM gives a hard decision. In contrast, the srRM yields higher mask values than the sRM, i.e., the srRM preserves more speech and noise energy than the sRM. To take the advantages of both RMs, the sRM, $Ms_{b,t}^{J^{sRM}}$, and srRM, $Ms_{b,t}^{J^{srRM}}$, are combined to form ssRM, $Ms_{b,t}^{J^{ssRM}}$, for the speech in Equation (31) and similarly, the ssRM, $Mr_{b,t}^{J^{ssRM}}$, for the speech RM is formed by combining the sRM, $Mr_{b,t}^{J^{sRM}}$, and srRM, $Mr_{b,t}^{J^{srRM}}$, in Equation (32) as follows:

$$Ms_{b,t}^{J^{ssRM}} = \alpha.Ms_{b,t}^{J^{sRM}} + (1-\alpha).Ms_{b,t}^{J^{srRM}} \tag{31}$$

$$Mr_{b,t}^{J^{ssRM}} = \alpha.Mr_{b,t}^{J^{sRM}} + (1-\alpha).Mr_{b,t}^{J^{srRM}} \tag{32}$$

Then, the de-framing process is applied to $Ms_{b,t}^{J^{ssRM}}$ in Equation (31) and $Mr_{b,t}^{J^{ssRM}}$ in Equation (32) to obtain a speech sequence, $ms_{b,t}^{J^{ssRM}}$, and a speech RM sequence, $mr_{b,t}^{J^{ssRM}}$, where both sequences have the same length as the original subband signal, $s_{b,t}^{J}$. So, we estimate the subband speech signal, $\hat{s}_{b,t}^{J}$, in two ways as:

$$\hat{s}_{b,t}^{J} = x_{b,t}^{J}.\times ms_{b,t}^{J^{ssRM}} \tag{33}$$

$$\hat{s}_{b,t}^{J} = x_{b,t}^{J}.\times mr_{b,t}^{J^{ssRM}} \tag{34}$$

5.4. Subband Power Normalization

A power normalization scheme is applied to the estimated subband signal, $\hat{s}_{b,t}^{J}$, in Equation (33) or (34) to obtain the power normalized subband signal, $\tilde{s}_{b,t}^{J}$, as:

$$\tilde{s}_{b,t}^{J} = \frac{\sigma_{b,t,c}}{\sigma_{b,t}}\hat{s}_{b,t}^{J} \tag{35}$$

where $\sigma_{b,t,c}$ represents the standard deviation of the clean speech utterance, which is calculated in the training phase, and $\sigma_{b,t}$ denotes the standard deviation of the estimated subband signal, $\hat{s}_{b,t}^{J}$. Finally, the enhanced speech signal, $\tilde{s}(t)$, can be obtained by applying the inverse DTCWT (IDTCWT) to the power normalized subband signal, $\tilde{s}_{b,t}^{J}$, in Equation (35).

6. Experiments and Discussion

Firstly, the experimental setup and evaluation techniques are introduced briefly. Secondly, the efficiency of our proposed method is diagnosed with STOI scores. Thirdly, the effects of the ARMA filter and ssRM are inspected. Fourthly, we compare the overall performance of our proposed method with the baseline (noisy data), STFT-NMF, and DWPT-NMF regarding HASQI and HASPI. Fifthly, the overall performance of the DTCWT-NMF with the previous three methods and the DNN-IRM are examined. Sixth, the unseen noise case of DTCWT-NMF with the unseen noise case of the baseline, STFT-NMF, DWPT-NMF, and DNN-IRM methods are compared concerning STOI and PESQ scores. Seventhly, we consider a multi-speaker dataset and compare our proposed method with the DNN-IRM and the DWPT-NMF regarding STOI and PESQ. Finally, we show the spectrograms of the clean speech, noisy speech, estimated speech through STFT-NMF, estimated speech through DWPT-NMF, estimated speech through DNN-IRM, and estimated speech through our proposed method.

6.1. Experimental Setup

For the experiment, an IEEE corpus [36] spoken by a single male speaker with 720 utterances at a sampling rate of 25 kHz is utilized to train, tune, and test our proposed system. We consider only 300 utterances for the training set by choosing randomly from 1 to 400, 100 utterances for the tuning set by selecting from 400 to 550, and another 100 utterances for the test set by choosing randomly from 551 to 720. Every utterance is downsampled at 16 kHz, and the length of these utterances is around 2 to 3 s. Eleven types of noises (e.g., stationary, non-stationary, quasi-stationary, and speech-shaped noises) are used for training, tuning, and testing purposes: Babble, birds, cafe, car, casino, factory, keyboard, machine gun, PC fan, speech-shaped noise (SSN), and street, which are drawn from different databases, such as the Aurora-2 database [37] and NOISEX-92 dataset [38]. Each noise is artificially added to the clean training, tuning, and test signals to generate the training, tuning and test noisy signals at five signal-to-noise ratios (SNRs) ranging from 10 dB to –10 dB with a 5 dB interval. Each noise length is around 30 s. Also, a random cut from the first 15 s of each noise is mixed with each training (tuning) utterance to generate the training (tuning) set. A random cut from the last 15 s of each noise is combined with each testing utterance to create the testing mixtures. Dividing the noise into two halves ensures that the testing noise segments are unseen during training. The applied magnitude spectrograms for STFT-NMF are computed by using a 1024 sample hamming window with 50% overlapping. The number of frames of the basis matrix is 40 for all the methods, the frame size is 1024 samples, and the frameshift is 20 samples followed in [24] for the

DWPT-NMF and DTCWT-NMF methods, respectively. Furthermore, the level of the wavelet is 3, the mother wavelet functions are Daubechies16 (db16) for DWPT/IDWPT and dtf2 for DTCWT/IDTCWT, and the iterations for training and testing are 100 and 50, respectively.

6.2. Evaluation Methods

We used the hearing aid's speech quality index (HASQI) [39], the hearing aid's speech perception index (HASPI) [40], the perceptual evaluation of speech quality (PESQ) [41], and the short-time objective intelligibility (STOI) [42] to evaluate the speech quality and intelligibility. Both of the HASQI and HASPI scores range from 0 to 1, and higher scores correspond to better sound quality and intelligibility, respectively. The PESQ is chosen as the objective quality test. It is widely used to assess the objective quality of speech signals. It gives scores in the range of −0.5 to 4.5, where higher scores correspond to better speech quality. STOI returns scores between 0 and 1, where higher STOI values imply better intelligibility. Besides, we calculated the performance of these enhancement methods (STFT-NMF, DWPT-NMF, DNN-IRM, and DTCWT-NMF) using the aforementioned metric scores by averaging over 100 test signals and 11 types of noises. The experiments 6.3 to 6.6 are handled by using the training set and the tuning set for the training stage and test stage, respectively. The best methods or parameters are then used to test the overall performance in experiment 6.7 by using the test set for the test stage.

6.3. The Efficiency of the Proposed System

In this subsection, we study the impact of the proposed method over the DWPT-NMF method with the ED and KL cost functions regarding STOI scores at different SNR levels. Figure 3 shows the performance comparison between the proposed method and the other methods, including the DWPT-NMF-ED, DWPT-NMF-KL, and DTCWT-NMF-ED methods, where the test signal is mixed with babble noise. Figure 3 shows that the proposed method, DTCWT-NMF-KL, has higher STOI values compared to the other methods except for the noise level of 5 dB. Therefore, we employed the DTCWT-NMF-KL method (referred to this paper as DTCWT-NMF) in the overall performance evaluation section.

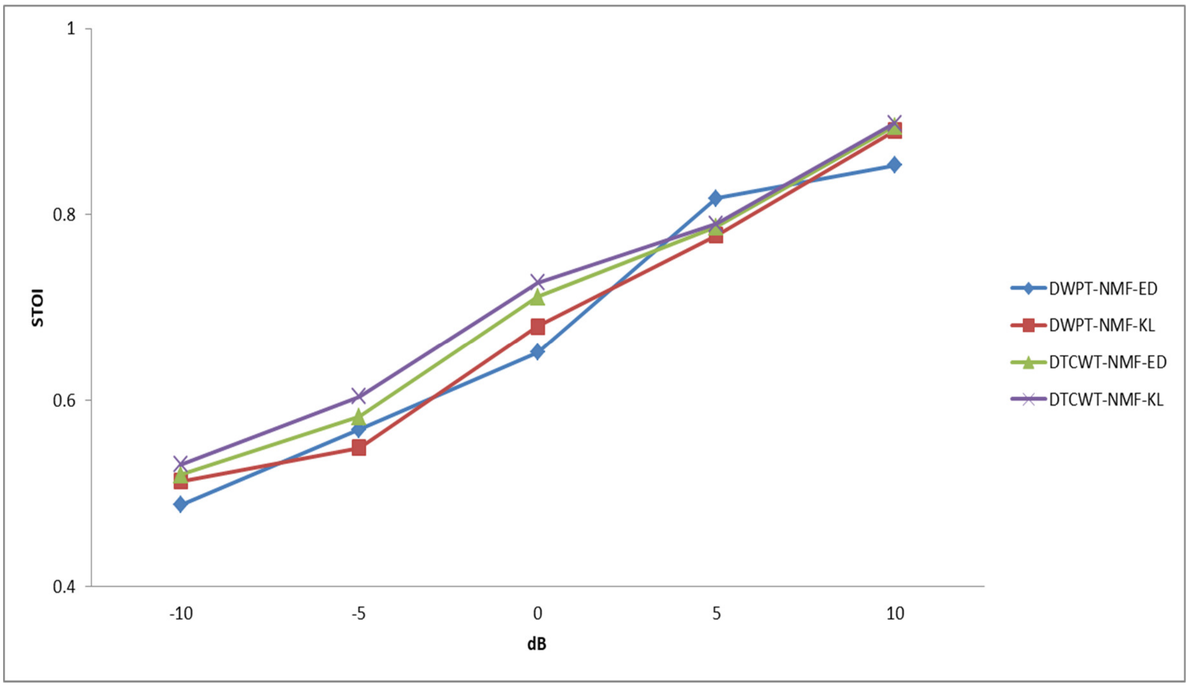


Figure 3. Comparison between DTCWT-NMF and DWPT-NMF methods with ED and KL.

6.4. The Effect of Speech and Noise Ratio Mask

We investigated the effect of speech and noise RMs that are learned jointly with the speech and noise training data considering four noises at five SNR levels as shown in Figure 4. According to Figure 4, we can achieve higher STOI values in all given noises and SNR levels based on speech and noise RMs. Also, we observed that the improvements of the STOI values in low SNR cases are gradually higher than high SNR cases. So, we exploited the speech and noise RMs in the overall performance evaluation section.

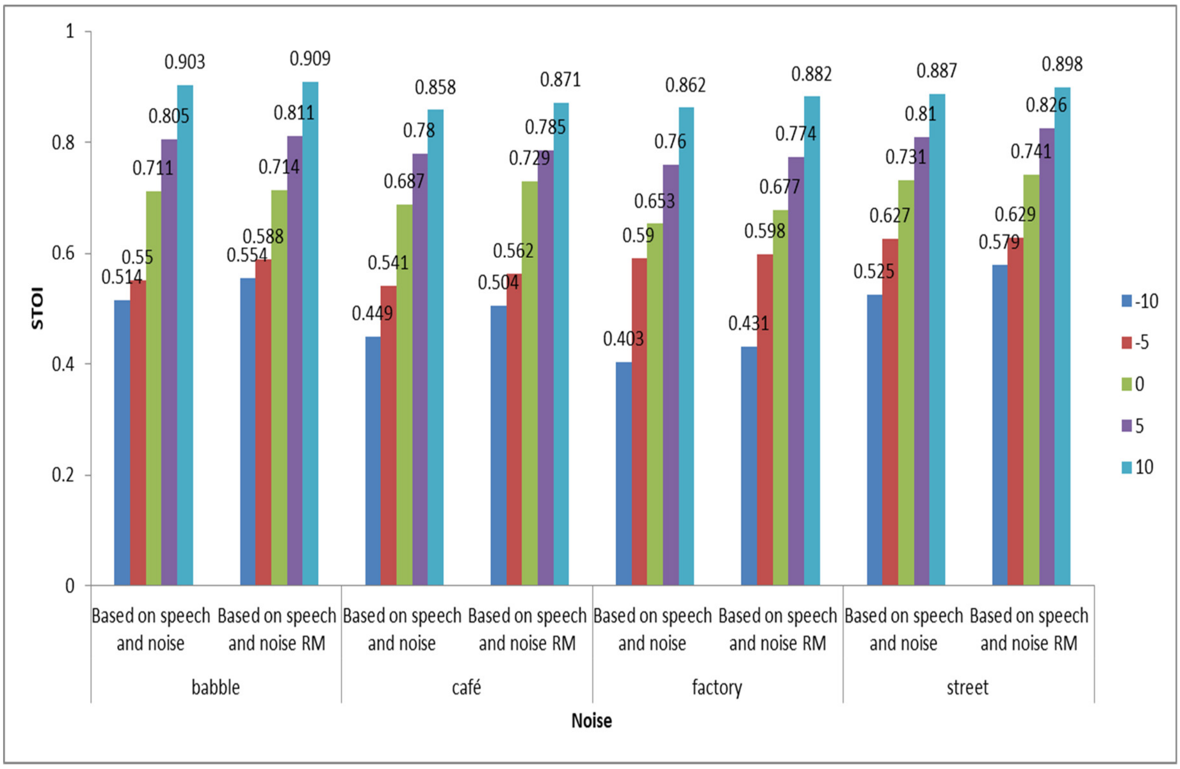


Figure 4. The effect of the speech and noise RM over speech and noise.

6.5. *The Effect of the ARMA Filter*

We observed and compared the outcomes of an ARMA filter in the system performances concerning STOI scores for four noises at five SNR cases. As we can see from Figure 5, it delivers a better impact on the system performance compared with the system performance without using the ARMA filter. So, we used the ARMA filter in the overall performance evaluation.

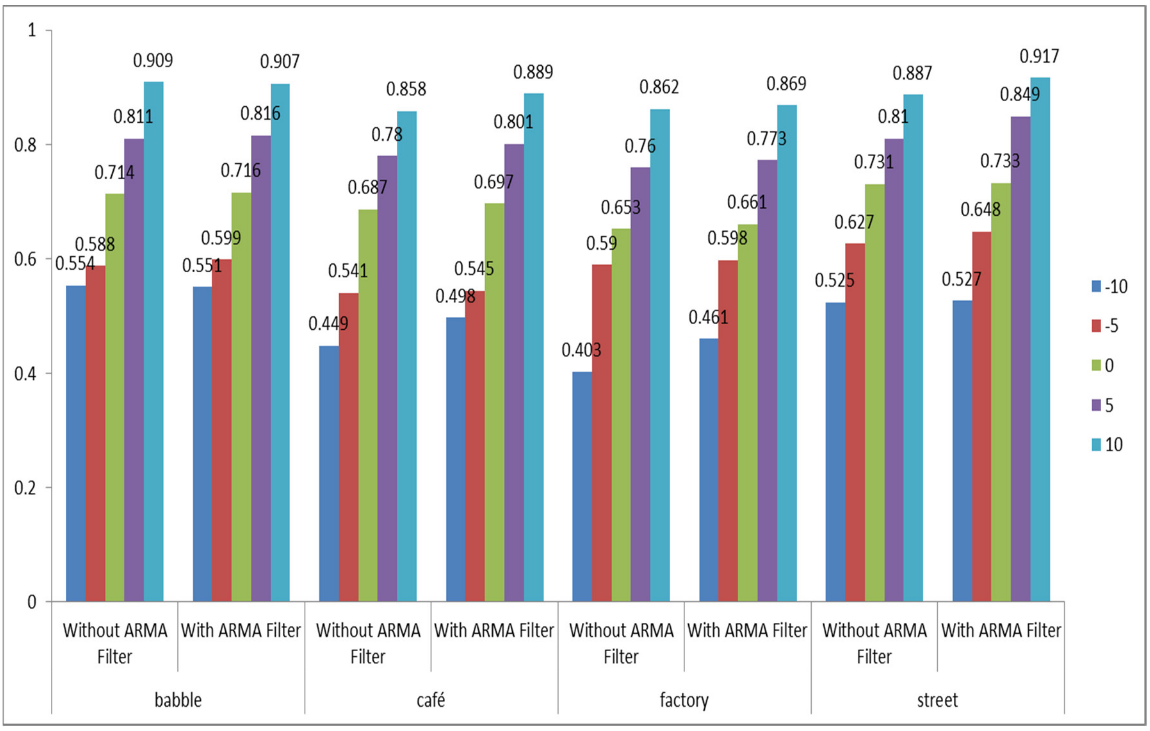


Figure 5. The impact of using the ARMA filter before decomposition via NMF.

6.6. Impact of the Proposed Subband Smooth Ratio Mask

Finally, we analyzed the impact of ssRM on STOI values at different SNR conditions. Figure 6 illustrates the ssRM at different α values. If we consider α = 0.8, it has a better STOI value in almost all SNR conditions than the other two α values as can be seen in Figure 6. So, the value of α is considered 0.8 in Equations (31) and (32).

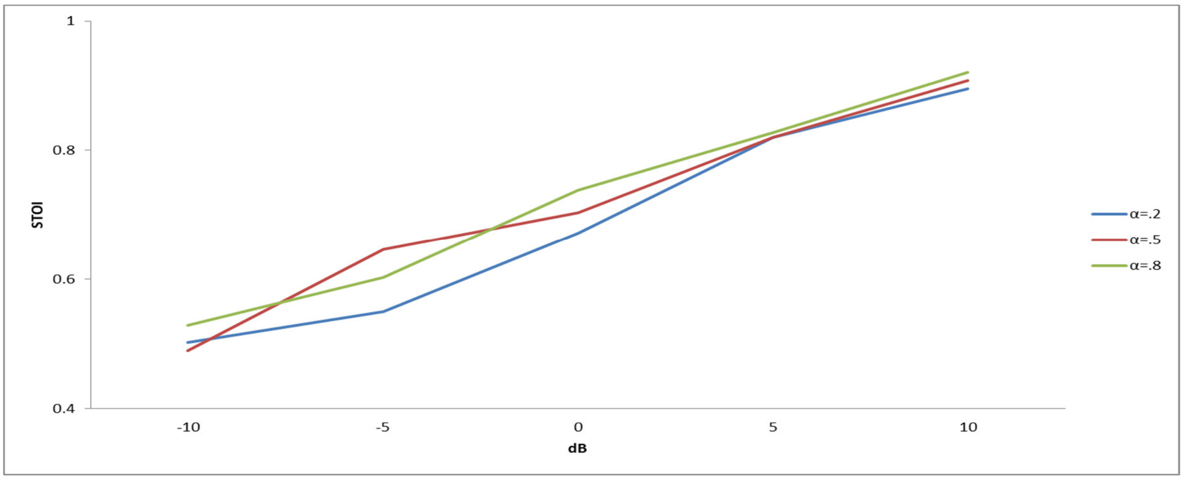


Figure 6. Effect of ssRM regarding the value of α.

6.7. Overall Performance Evaluation

Based on the HASQI and HASPI metrics, we compared the performance of the proposed method with the baseline (unprocessed noisy speech), STFT-NMF, and DWPT-NMF methods. Tables 1 and 2 show the experimental results of five SNR cases. According to Tables 1 and 2, we confirmed that the proposed method provides better HASQI and HASPI scores than the STFT-NMF, DWPT-NMF, and baseline methods for all SNR cases. For this consequence, we presume that the DTCWT-NMF method improves the performance significantly concerning the quality and intelligibility of the speech signal. We also find that the HASQI scores are gradually improved from high to low SNR cases of the DTCWT-NMF method, whereas the HASPI scores are abruptly improved in low SNR cases than high SNR. Ultimately, we speculate that the proposed method works well in low SNR cases based on HASPI values.

Table 1. The HASQI results in five SNR cases.

Method	–10	–5	0	5	10
Baseline	0.058	0.079	0.110	0.143	0.180
STFT-NMF	0.053	0.086	0.139	0.183	0.204
DWPT-NMF	0.067	0.127	0.216	0.298	0.359
DTCWT-NMF	**0.121**	**0.216**	**0.354**	**0.505**	**0.633**

Table 2. The HASPI results in five SNR cases.

Method	–10	–5	0	5	10
Baseline	0.307	0.398	0.571	0.731	0.848
STFT-NMF	0.269	0.468	0.763	0.933	0.970
DWPT-NMF	0.333	0.596	0.886	0.978	0.993
DTCWT-NMF	**0.547**	**0.787**	**0.970**	**0.997**	**0.999**

Tables 3 and 4 list the PESQ and STOI values of different methods, including the baseline, STFT-NMF, DWPT-NMF, DNN-IRM, and DTCWT-NMF methods, under five SNR conditions. From Table 3, it can be seen that the DTCWT-NMF yields better PESQ values than the baseline, STFT-NMF, and DWPT-NMF methods in all SNR conditions. Contrarily, the DTCWT-NMF method outperforms the DNN-IRM for PESQ values under low SNR conditions. Similar performance trends can be observed from Table 4 for STOI values.

Table 3. The PESQ values under five SNR cases.

Method	–10	–5	0	5	10
Baseline	1.435	1.539	1.709	1.899	2.035
STFT-NMF	1.463	1.559	1.733	1.930	2.053
DWPT-NMF	1.493	1.659	1.851	2.041	2.216
DNN-IRM	1.520	1.716	**2.031**	**2.465**	**2.761**
DTCWT-NMF	**1.551**	**1.725**	2.006	2.298	2.614

Table 4. The STOI results in five SNR cases.

Method	–10	–5	0	5	10
Baseline	0.533	0.591	0.652	0.717	0.784
STFT-NMF	0.451	0.504	0.578	0.632	0.677
DWPT-NMF	0.522	0.623	0.717	0.787	0.835
DNN-IRM	0.547	0.654	**0.757**	**0.844**	**0.905**
DTCWT-NMF	**0.572**	**0.660**	**0.757**	0.841	0.903

In the unseen noise case, we considered five types of noises: Babble, birds, cafe, car, and casino for training and six types of noises: Factory, keyboard, machine gun, pc fan, speech-shaped noise

(SSN), and street for testing. Moreover, we calculated the performance of the baseline, U-STFT-NMF, U-DWPT-NMF, U-DNN-IRM, and U-DTCWT-NMF methods using the STOI and PESQ metrics by averaging over 100 test signals and six types of noises. It is noted that the prefix, U, of these methods indicates the unseen noise case of the STFT-NMF, DWPT-NMF, DNN-IRM, and DTCWT-NMF methods. Table 5 compares the performance of the U-DTCWT-NMF method with the other three techniques, including U-STFT-NMF, U-DNN-IRM, and U-DWPT-NMF, and the baseline at five SNR conditions. By observing Table 5, it can be seen that the U-DTCWT-NMF method outperforms the U-STFT-NMF and U-DWPT-NMF regarding STOI and PESQ in all SNR cases, whereas the U-DTCWT-NMF outperforms the U-DNN-IRM method in low SNR cases. Thus, it confirms our previous conclusion that the performance of the proposed method is better than the other three methods, including STFT-NMF, DWPT-NMF, and DNN-IRM, and the baseline in low SNR cases.

Table 5. Comparison between the U-DTCWT-NMF and the other three unseen noise case methods.

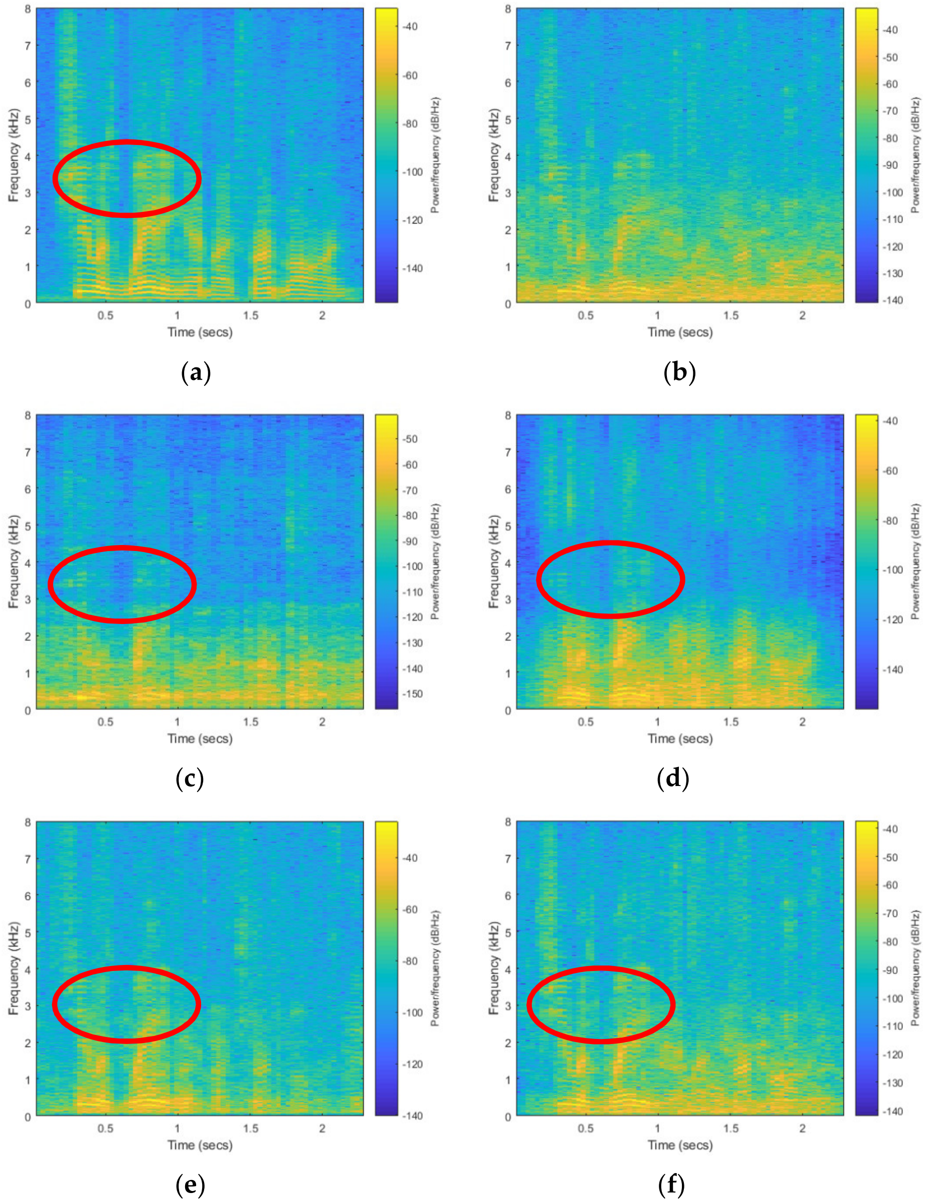
Method	**-10**	**-5**	**0**	**5**	**10**
	STOI/PESQ	**STOI/PESQ**	**STOI/PESQ**	**STOI/PESQ**	**STOI/PESQ**
Baseline	0.551/1.352	0.641/1.646	0.734/1.901	0.829/2.256	0.902/2.540
U-STFT-NMF	0.477/1.349	0.545/1.553	0.609/1.775	0.682/1.987	0.728/2.188
U-DWPT-NMF	0.534/1.407	0.623/1.666	0.716/1.962	0.789/2.307	0.847/2.584
U-DNN-IRM	0.542/1.427	0.652/1.681	0.764/1.970	**0.855/2.343**	**0.917/2.692**
U-DTCWT-NMF	**0.569/1.437**	**0.667/1.707**	**0.765/1.973**	0.849/2.323	0.909/2.642

In the multi-speaker case, we collected the speech signals and non-stationary noises from the GRID audio-visual corpus [43] and the NOISEX-92 corpus [38], respectively. The noise types include factory floor noise, tank noise, military vehicle noise, and speech babble. Ten male and ten female speakers' utterances were used to form an experimental group for each noise type under each input SNR of –10 dB, –5 dB, 0 dB, 5 dB, and 10 dB. In each group, we used the speech and noise segments of 60 seconds for training and the mixture segment of 10 seconds for testing. Meanwhile, the results of 20 speakers in each group were averaged for every noise type under one input SNR. Table 6 comprises the comparison of the performances for the DWPT-NMF, DNN-IRM, and DTCWT-NMF regarding STOI and PESQ metrics under five SNR cases. As shown in Table 6, the proposed method outperforms the other two methods at low SNR conditions concerning STOI and PESQ.

Table 6. Performance comparison among three methods across five SNR cases.

Method	**–10**	**–5**	**0**	**5**	**10**
	STOI/PESQ	**STOI/PESQ**	**STOI/PESQ**	**STOI/PESQ**	**STOI/PESQ**
DWPT-NMF	0.495/1.37	0.572/1.591	0.665/1.782	0.724/2.041	0.796/2.216
DNN-IRM	0.522/1.481	0.614/1.686	**0.731/1.985**	**0.832/2.381**	**0.912/2.653**
DTCWT-NMF	**0.531/1.551**	**0.632/1.702**	0.729/1.981	0.829/2.195	0.892/2.46

The noisy spectrogram is displayed in Figure 7b, which is a mixture of clean speech as illustrated in Figure 7a and babble noise at 0 dB SNR. The spectrograms of the enhanced speech signals applying the STFT-NMF, DWPT-NMF, DNN-IRM, and DTCWT-NMF methods are visualized in Figures 7c–f, respectively. These figures depict that the STFT-NMF and DWPT-NMF methods show higher speech distortion than the DTCWT-NMT method when the frequency is high. Thus, the proposed method successfully eliminates the background noise from the mixed signal even when the noise level is high and concurrently obtains a better estimation of the speech signal with fewer distortions (red circles in the figures indicate the location).



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 7. Spectrogram of (**a**) clean speech, (**b**) noisy, (**c**) STFT-NMF, (**d**) DWPT-NMF, (**e**) DNN-IRM, and (**f**) DTCWT-NMF where the x-axis corresponds to the time in seconds and the y-axis corresponds to the frequency in kHz.

7. Conclusion

In this study, we proposed a novel speech enhancement method that adheres to the DTCWT and NMF with the KL cost function to achieve better performance of speech signal from noisy signal. The main focuses of our research were the development of the wavelet-based speech enhancement method over the STFT based method, estimation of RMs through the joint learning process, application of an ARMA filter to achieve better decomposition results in NMF, and the adoption of ssRM. Furthermore, better results were achieved by considering small training data, less iterations, and maintaining redundancy at an acceptable level. Systematic evaluation using objective metrics indicated that the proposed method should improve speech quality and intelligibility in a wide range of noisy conditions. The interesting point is that the proposed method works better than the DNN-IRM method concerning STOI and PESQ metrics at low SNR conditions, since the DTCWT decomposes the time-domain signal into a set of subband signals and obtains a good time-frequency resolution (i.e., good time-frequency resolution means high-frequency components of a signal have good time resolutions and low-frequency components of a signal have good frequency resolutions).

Subsequently, the noise can be properly eliminated from the noisy signal via NMF. In the case of unseen noise, the proposed method significantly outperforms the baseline, STFT-NMF, DWPT-NMF, and DNN-IRM methods in low SNR conditions as shown in the experimental results. In the future, we intend to consider better mother wavelets to improve this algorithm and a deep neural network (DNN) for secondary estimation.

Author Contributions: Conceptualization, M.S.I. and T.H.A.M.; Methodology, M.S.I. and W.U.K.; Software, M.S.I.; Writing—original draft preparation, M.S.I.; Writing—review and editing, T.H.A.M. and Z.Y; Supervision, Z.Y.

Funding: This work is supported by the National Natural Science Foundation of China (No. 61671418) and the Advanced Research Fund of the University of Science and Technology of China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boll, S.F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120.
2. Wang, J.; Liu, H.; Zheng, C.; Li, X. Spectral subtraction based on two-stage spectral estimation and modified cepstrum thresholding. *Appl. Acoust.* **2013**, *19*, 450–458.
3. Mcaulay, R.; Malpass, M. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *65*, 137–145.
4. Lotter, T.; Vary, P. Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP J. Appl. Signal Process.* **2005**, *7*, 1110–1126.
5. Ephraim, Y.; Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 1109–1121.
6. Saleem, N.; Khattak, M.I.; Shafi, M. Unsupervised speech enhancement in low SNR environments via sparseness and temporal gradient regularization. *Appl. Acoust.* **2018**, *141*, 333–347.
7. Scalart, P.; Filho, J.V. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, USA, 9 May 1996; Volume 2, pp. 629–632.
8. Djendi, M.; Bendoumia, R. Improved subband-forward algorithm for acoustic noise reduction and speech quality enhancement. *Appl. Soft Comput.* **2016**, *42*, 132–143.
9. Ephraim, Y.; Trees, H.L.V. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266.
10. Narayanan, A.; Wang, D.L. Ideal ratio masks estimation using deep neural networks for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7092–7096.
11. Kang, T.G.; Kwon, K.; Shin, J.W.; Kim, N.S. NMF-based target source separation using deep neural network. *IEEE Signal Process. Lett.* **2015**, *22*, 229–233.
12. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Speech enhancement based on deep denoising autoencoder. In Proceedings of the INTERSPEECH, Lyon, France, 25–29 August 2013; pp. 436–440.
13. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322.
14. Lee, H.; Battle, A.; Raina, R.; Ng, A.Y. Efficient sparse coding algorithms. *Adv. Neural Inf. Process. Syst.* **2007**, *19*, 801–808.
15. Chen, Z.; Ellis, D. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
16. He, Y.; Sun, G.; Han, J. Spectrum enhancement with sparse coding for robust speech recognition. *Digit. Signal Process.* **2015**, *43*, 59–70.
17. Luo, Y.; Bao, G.; Xu, Y.; Ye, Z. Supervised monaural speech enhancement using complementary joint sparse representations. *IEEE Signal Process. Lett.* **2016**, *23*, 237-241.
18. Wilson, K.W.; Raj, B.; Smaragdis, P.; Divakaran, A. Speech denoising using nonnegative matrix factorization with priors. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4029–4032.

19. Lee, S.; Han, D.K.; Ko, H. Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities. *Appl. Acoust.* **2017**, *117*, 257–262.
20. Mowlaee, P.; Saeidi, R.; Stilanou, Y. Phase importance in speech processing applications. In Proceedings of the INTERSPEECH, Singapore, 14–18 September 2014; pp. 1623–1627.
21. Ghanbari, Y.; Karami-Mollaei, M.R. A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. *Speech Commun.* **2006**, *48*, 927–940.
22. Ghribi; Djendi, M.; Berkani, D. A wavelet-based forward BSS algorithm for acoustic noise reduction and speech enhancement. *Appl. Acoust.* **2016**, *105*, 55–66.
23. Jung, S.; Kwon, Y.; Yang, S. Speech enhancement by wavelet packet transform with best fitting regression line in various noise environments. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Toulouse, France, 14–19 May 2006; Volume 1, pp. 14–19.
24. Wang, S.-S.; Chern, A.; Tsao, Y.; Hung, J.-W.; Lu, X.; Lai, Y.-H.; Su, B. Wavelet speech enhancement based on nonnegative matrix factorization. *IEEE Signal Process. Lett.* **2016**, *23*, 1101 – 1105.
25. Messaoud, M.A.B.; Bouzid, A.; Ellouze, N. Speech enhancement based on wavelet packet of an improved principal component analysis. *Comput. Speech Lang.* **2016**, *35*, 58–72.
26. Mavaddaty, S.; Ahadi, S.M.; Seyedin, S. Speech enhancement using sparse dictionary learning in wavelet packet transform domain. *Comput. Speech Lang.* **2017**, *44*, 22–47.
27. Mortazavi, S.H.; Shahrtash, S.M. Comparing Denoising Performance of DWT, DWPT, SWT and DT-CWT for Partial Discharge Signals. In Proceedings of the 43rd International Universities Power Engineering Conference, Padova, Italy, 1–4 September 2008; pp. 1–6.
28. Williamson, D.S.; Wang, Y.; Wang, D.L. Reconstruction techniques for improving the perceptual quality of binary masked speech. *J. Acoust. Soc. Am.* **2014**, *136*, 892–902.
29. Wang, Y.; Narayanan, A.; Wang, D.L. On training targets for supervised speech separation. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858.
30. Williamson, D.S.; Wang, Y.; Wang, D.L. Complex ratio masking for monaural speech separation. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 483–493.
31. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **2001**, *13*, 556–562.
32. Kingsbury, N.G. The dual-tree complex wavelet transform: A new efficient tool for image restoration and enhancement. In Proceedings of the 9th European Signal Process Conference, EUSIPCO 1998, Rhodes, Greece, 8–11 September 1998; pp. 319–322.
33. Selenick, I.W.; Baraniuk, R.G.; Kingsbury, N.G. The dual-tree complex wavelet transforms. *IEEE Signal Process. Mag.* **2005**, *22*, 123–151.
34. Mohammadiha, N.; Taghia, J.; Leijon, A. Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions. In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Kyoto, Japan, 25–30 March 2012; pp. 4561–4564.
35. Chen, C.-P.; Bilmes, J. MVA Processing of Speech Features. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 257–270.
36. Rothauser, E.H. IEEE recommended practice for speech and quality measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 225–246.
37. Hirsch, H.G.; Pearce, D. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In Proceedings of the ISCA Tutorial and Research Workshop, ISCA ITRWASR, Paris, France, 18–20 September 2000; pp. 181–188.
38. Varga, A.; Steeneken, H.J.M. 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.* **1993**, *12*, 247–251.
39. Kates, J.M.; Arehart, K.H. The hearing-aid speech quality index (HASQI). *J. Audio Eng. Soc.* **2010**, *58*, 363–381.
40. Kates, J.M.; Arehart, K.H. The hearing-aid speech perception index (HASPI). *Speech Commun.* **2014**, *65*, 75–93.
41. Rix, A.; Beerends, J.; Hollier, M.; Hekstra, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.

42. Tall, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136.

43. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424.