Applied Acoustics 174 (2021) 107732

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain

Longting Xu [a], Zhilin Wei [a], Syed Faham Ali Zaidi [a], Bo Ren [b], Jichen Yang [c,*]

[a] College of Information Science and Technology, Donghua University, Shanghai, China
[b] Microsoft Search Technology Center Asia, Suzhou, China
[c] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 26 February 2020
Received in revised form 11 September 2020
Accepted 9 October 2020
Available online 15 November 2020

Keywords:
Constant-Q transform
Spectrogram
Speech enhancement
Additive noise
NMF
SNMF

ABSTRACT

The utterance can be easily affected by additive noise in a real environment. To decrease the additive noise, the noisy speech can be enhanced based on the spectrogram following with Nonnegative Matrix Factorization (NMF) and sparse NMF(SNMF) algorithm. More information can be obtained at a high sampling rate. The range of objective human vocal organs is limited to a low-frequency value compared to the high sampling rate; thus, higher resolution is required to describe the low frequencies. Traditional spectrogram based on short-time Fourier transform (STFT) may lack frequency resolution at lower frequencies. To this end, we propose to use a constant Q transform (CQT) in this paper, which can give high resolution for the low frequencies. The backend algorithm remains the NMF/SNMF algorithm. We evaluate the proposed method with the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI). The experimental results show that our proposed method shows better enhancement ability compared to the STFT baseline at low Signal to Noise Ratio (SNR).

1. Introduction[1]

Speech signals are often degraded by the presence of noise, arising from the transmission equipment or the surrounding environment. There are a variety of speech enhancement algorithms, and most of them can be categorized into two types: unsupervised-based and supervised-based. With the unsupervised-based methods, such as Wiener and Kalman filter [1], estimators based on optimally modified log-spectral amplitude (OM-LSA) [2] and minimum mean-square error short-time spectral amplitude (MMSE STSA) [3]. In these methods, the clean speech is computed from the estimated parameters of the model, and the noise type or the speaker identity does not need to be trained prior. However, supervised methods usually require prior information. The prior information could be obtained with dictionary learning (DL) [4,5] approaches. The most commonly used sparse representation algorithms based on dictionary learning includes: K-Singular Value Decomposition (K-SVD) [6], Probability Latent Component Analysis (PLCA) [7], Non-negative Matrix Factorization (NMF) [8,9]. K-SVD is a dictionary training algorithm extended from the K-means clustering method. PLCA algorithms and NMF algorithms have similar theories and processing methods. NMF generally gives better performance. At the same time, by combining with the non-negative nature of speech itself, NMF is widely used in the research of speech enhancement algorithms.

Due to the high factorization efficiency of the NMF, it is widely used in source separation and audio signal processing [10,11]. Several attempts have been investigated to enhance speech with NMF [12,13]. In [12], it aimed to train the dictionary with the NMF and adaptively update the noise dictionary for speech enhancement. In [13], an adaptive denoise method for sound event detection based on NMF is presented. Jonathan Le Roux et al. proposed the sparse NMF (SNMF) [20] in 2015. The SNMF is an improved NMF, which directly redefines the objective function and uses re-normalization in updating matrix. The input spectrum of the NMF is usually obtained by the well-known short-time Fourier transform (STFT). With the STFT method, the denoise performance is good, but the STFT approach may lack frequency resolution at lower frequencies [14]. While the range of objective human vocal organs is limited to a low-frequency value, which is usually lower than 4 kHz. More information can be obtained at a high sampling rate, and in this paper, the sampling rate is 24 kHz of the corpus. To describe the human speaker's speech more specifically, higher resolution is required to describe the low frequencies. To this

* Corresponding author.
E-mail address: nisonyoung@163.com (J. Yang).

[1] This work has been supported by Shanghai Sailing Program (No.19YF1402000), the Fundamental Research Funds for the Central Universities (No.2232019D3-52), and the Initial Research Funds for Young Teachers of Donghua University.

https://doi.org/10.1016/j.apacoust.2020.107732
0003-682X/

end, we propose to use a constant Q transform (CQT) for speech enhancement in this paper. The CQT is a transformation with a constant ratio of center frequency to bandwidth. In order to make the constant Q cross the entire spectrum, the CQT has a higher frequency resolution at lower frequencies while provide a higher temporal resolution at higher frequencies. In [15], CQT was used to solve the problem of musical tone transfer. In [16], CQT analyzed the spectrum of music signals. In recent years, CQT has been increasingly used in the field of speech signal processing. For example, in [17,18], one important countermeasure of spoofing attack detection derived from CQT was proposed, and it was named as constant-Q cepstral coefficients (CQCC).

Based on the CQT and NMF/SNMF, we propose a new method for speech enhancement. To decrease the additive noise in the speech, the proposed steps for speech enhancement are as follows: in the off-line phase, first, parameters matrix V_S and V_n of pure speech signals $s(t)$ and noise signals $n(t)$ are obtained based on the CQT, which mainly concludes the amplitude matrix. Second, we train pure speech dictionary W_s and noise dictionary W_n with the NMF method, respectively, and constitute a dual dictionary$[W_s\ W_n]$. In the on-line phase, first, we extract the amplitude matrix V of noise-corrupted speech $y(t)$ based on the CQT. Then we use the speech dictionary to obtain the reconstructed speech amplitude matrix$V$. Finally, we add the phase of noise-corrupted speech for the enhanced speech, then obtain the enhanced speech time-domain signal $s(t)$ with inverse constant Q transform (ICQT).

In the community of speech signal processing, we often need to consider contextual information. Some speech features are cascaded and may span several frames. Therefore, multiple-frame exemplars were used in the speech dictionary to more accurately estimate the activation matrix [19]. In this paper, we use a total of 5 frames which include left two frames, right two frames, and the current frame as a multi-frame input.

The rest of the paper is organized as follows. SectionⅡ reviews the methodology of NMF, SNMF and corresponding baseline. Section III reviews the knowledge of CQT in detail and proposes a specific speech enhancement method based on NMF-CQT. Section Ⅳ evaluates our proposed method under Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) criteria, and Section Ⅴ concludes the paper.

2. Speech enhancement based on NMF/SNMF-STFT algorithm

2.1. Basic concepts of the NMF

NMF was introduced by Lee and Seung [8] which projects a non-negative matrix onto space spanned by its nonnegative basis vectors. A non-negative matrix V is decomposed into two matrices W and H. The goal of the NMF algorithm is to find the WH closest to V. $V \approx W \times H$ **,** it transforms the matrix factorization algorithm into the following optimization problem, that is, an optimization problem that minimizes the Euclidean distance between two matrices:

$$min\|V - V^{'}\|^2 = \sum_{ij}\left(V_{ij} - V^{'}_{ij}\right)^2 \# \tag{1}$$

where V is the original matrix and V' is the matrix to be updated, that is $V^{'} = W \times H$, where W is the basis matrix, H is the coefficient matrix. The multiplication update rules are as follows:For the loss function of Euclidean distance:

$$W_{ij}^{k+1} = W_{ij}^{k}\frac{\left(\left(H^K\right)^T * V\right)_{ij}}{\left(\left(H^K\right)^T * W^k * H^k\right)_{ij}} \# \tag{2}$$

$$H_{ij}^{k+1} = H_{ij}^{k}\frac{\left(\left(W^K\right)^T * V\right)_{ij}}{\left(\left(W^K\right)^T * W^k * H^k\right)_{ij}} \# \tag{3}$$

For the Kullback-Leibler (KL) divergence loss function:

$$W_{i,k} = W_{i,k}\frac{\sum_u H_{k,u}V_{i,u}/(WH)_{i,u}}{\sum_v H_{k,v}} \# \tag{4}$$

$$H_{i,k} = H_{i,k}\frac{\sum_u W_{u,k}V_{i,u}/(WH)_{u,j}}{\sum_v W_{v,k}} \# \tag{5}$$

Once matrices W and H are initialized with random non-negative values, the multiplicative update rules can preserve their non-negativity during iteration.

2.2. Basic concepts of the SNMF

Jonathan Le Roux et al. proposed the SNMF (sparsity NMF) [20] in 2015. Compared with NMF, SNMF achieves more appropriate sparsity. And SNMF is to re-normalize the W matrix after each multiplication update to get the normalized version W, but not rescale H accordingly. The calculation with a β-divergence equation of the gradient [29] of $D_\beta(M|WH)$ with W_i is as follows:

$$\nabla_{W_i} D_\beta(M|WH) = \left((WH)^{\beta-2} \otimes (WH - M)\right)H_i^T \# \tag{6}$$

where the $\otimes$ product and the addition of exponents are treated as element modes. For more clarity, we use the notation $\Lambda = WH$. The calculation method of the gradient of $D_\beta\left(M|WH\right)$ with W_i is as follows:

$$\nabla_{W_i} D_\beta(M|\Lambda) = \frac{1}{\|W_i\|}\left(Id - W_i\ W_i^T\left(\Lambda^{\beta-2} \otimes (\Lambda - M)\right)\right)H_i^T \# \tag{7}$$

Therefore, the multiplication update equation of the W matrix is as follows:

$$W \leftarrow W \otimes \frac{\left(\Lambda^{\beta-2} \otimes M\right)H^T + W \otimes \left(11^T\left(W \otimes \left(\Lambda^{\beta-1}H^T\right)\right)\right)}{\Lambda^{\beta-1}H^T + W \otimes \left(11^T\left(W \otimes \left(\left(\Lambda^{\beta-2} \otimes M\right)H^T\right)\right)\right)} \# \tag{8}$$

where 1 is a column vector with all elements equal to 1.

2.3. Speech enhancement based on NMF/SNMF in STFT domain

The magnitude spectrograms of the noise-corrupted speech $y(t)$ are the sum of the speech magnitude spectrograms S and noise spectrograms N. Based on NMF/SNMF, the approximate posterior distributions of the speech basis matrix W_s, coefficient matrix H_s and noise basis matrix W_n, coefficient matrix H_n are obtained by the training data. Therefore, the basis matrix and coefficient matrix for speech denoising is founded by the tandem connection of W_s, W_n, H_s and H_n.

$$W = [W_s\ W_n] \# \tag{9}$$

$$H = [H_s; H_n] \# \tag{10}$$

This is the key to sparse decomposition of dictionary learning [21]. The speech components in noisy corrupted speech are coherent to the pure speech dictionary, and the noise part is coherent to the noise dictionary.

In a supervised framework, the W matrices of clean speech and noise, denoted as W_s and W_n respectively, are obtained first during the training stage. The NMF/SNMF decomposition for the denoising takes the following form [22].

$$V \approx WH = [W_s W_n]\begin{bmatrix} H_s \\ H_n \end{bmatrix} \# \tag{11}$$

The pure speech spectrogram can be separated from the noise and can be estimated by only using the pure speech counterpart W_s, H_s. that is.

$$\widehat{V_s} \approx \frac{W_s H_s}{W_s H_s + W_n H_n} \odot V \# \tag{12}$$

where the product and division are carried out in an element-wise fashion. Multiplicative update rules for the activation matrices of NMF algorithm are derived, as follows:

$$H_c \leftarrow H_c \odot \left(W_c^T \frac{V}{WH}\right) / \left(W_c^T 1\right) \quad where\ c = s, n \# \tag{13}$$

Multiplicative update rules for the activation matrices of SNMF algorithm are derived, as follows:

$$H \leftarrow H \otimes \frac{W^T \left(M \otimes \Lambda^{\beta-2}\right)}{W^T \Lambda^{\beta-1} + \mu} \# \tag{14}$$

where V is the matrix of noisy speech, and the updated coefficient matrix H can be obtained according to multiplicative update rules. We retain the coefficient matrix H_s of the pure speech part, and then multiply it with the prior dictionary W_s to obtain the pure speech amplitude matrix $V^{'}$. With the amplitude parameter matrix $V^{'}$ and the phase of noise-corrupted speech, we can obtain the enhanced speech time-domain signal with the ISTFT.

The speech spectrum obtained by the STFT may lack frequency resolution at lower frequencies and lack temporal resolution at higher frequencies. In contrast, the CQT can produce a higher frequency resolution at lower frequencies while providing a higher temporal resolution at higher frequencies. Therefore, to improve the auditory effect of speech, this paper proposes the speech enhancement can be conducted in the Constant-Q frequency domain with the NMF and SNMF algorithm.

3. Speech Enhancement Based on NMF/SNMF-CQT algorithm

3.1. Basic concepts of the CQT

The CQT approach was proposed by Brown and Pluckette in 1991 and successfully applied in the field of music signal processing [23]. In this paper, the NMF and SNMF algorithm based on the CQT is proposed to decrease the additive noise. The CQT is a transformation with a constant center frequency-bandwidth ratio. The frequency band of the exponential distribution after the CQT corresponds to the scale frequency of the music.

We see the essence of CQT from the calculation process. Center frequency to bandwidth ratio is a constant value Q

$$Q = \frac{f_k}{\delta f_k} \# \tag{15}$$

$$f_k = 2^{k/\beta} f_{min} \# \tag{16}$$

and

$$f_{k+1} - f_k = f_k \frac{1}{2^{1/\beta} - 1} \# \tag{17}$$

where f_{min} is the center frequency of the lowest-frequency bin, f_k represents the frequency of the k-th component, β is the number of frequency bin contained in an octave, such as β = 36, which means that there are 36 frequency bins in each octave.

Let δ_f represent the frequency bandwidth at frequency f, which can also be called frequency resolution. In the CQT, the number of bins per octave (B) is related to the fidelity factor (Q) by Equation (15) [24].

$$Q = \frac{f}{\delta_f} = \frac{1}{2^{1/\beta} - 1} \# \tag{18}$$

From this equation, we know that Q is only related to β. Next, we assume that N_k is the window length that changes with frequency, and f_s represents the sampling frequency

$$N_k = \frac{f_s}{\delta_{f_k}} = \frac{f_s Q}{f_k} \# \tag{19}$$

The CQT transform uses different window widths to obtain different frequency resolutions, so that the frequency amplitude of each semitone can be obtained. The k-th halftone frequency component of the N-th frame in the CQT can be expressed as:

$$X_n^{cqt}(k) = \frac{1}{N} \sum_{m=0}^{N_k - 1} x(m) w_{N_k}(m) e^{-j2\pi m Q / N_k} \# \tag{20}$$

where $x(m)$ is a time-domain signal, w_{N_k} is a window function, and X_n^{cqt} is a converted spectral parameter.

The CQT is the ratio of the center frequency to the bandwidth is a constant Q, which means that the length of the time-domain data required to calculate the high-band spectrum is shorter than that of the low-band spectrum [25]. In other words, the CQT has higher time-domain resolution and lower frequency resolution in the high-frequency band, it also has higher frequency domain resolution and lower time-domain resolution in the low-frequency band [26].

4. Proposed speech enhancement approach in the CQT domain

The work of section Ⅱ has proved the effectiveness of speech enhancement based on NMF in the STFT frequency domain. However, the converted spectrum by STFT is uniform in each frequency band. Due to the fact that, the human speech concentrates on the low-frequency band, and the CQT can produce a higher frequency resolution at lower frequencies. Therefore, in order to improve the auditory quality of speech, we propose speech enhancement in the CQT frequency domain. The process of speech enhancement is mainly divided into two phases: off-line phase, on-line phase.

4.1. Off-line phase

First, time-domain pure speech $s(t)$ and nosie $n(t)$ are converted into frequency-domain signals by CQT, respectively. On the basis of CQT, we can obtain a pure speech amplitude matrix V_S and noise amplitude matrix V_n. Second, the magnitude parameter matrices V_S and V_n are decomposed into base matrix W and parameter matrix H based on the NMF or SNMF algorithm. We normalize W and re-scale H after the update of W and H. This paper uses KL divergence to update the matrix. The update rules are equation (4) and equation (5).

A large amount of pure speech and noise are trained to obtain a speech dictionary W_s and a noise dictionary W_n. The whole basis matrix and coefficient matrix for speech denoising is founded by the tandem connection of W_s and W_n.

$$W = [W_s W_n] \# \tag{21}$$

4.2. On-line phase

To decrease the noise from noisy speech, we first obtain the amplitude spectrum V and phase spectrum of noisy-corrupted speech $y(t)$ with CQT. The pure speech spectrogram can then be separated from the noise and can be estimated by only using the pure speech counterpart W_s, H_s as shown in Equation (12).

$$\widehat{V_s} \approx \frac{W_s H_s}{W_s H_s + W_n H_n} \odot V \# \tag{22}$$

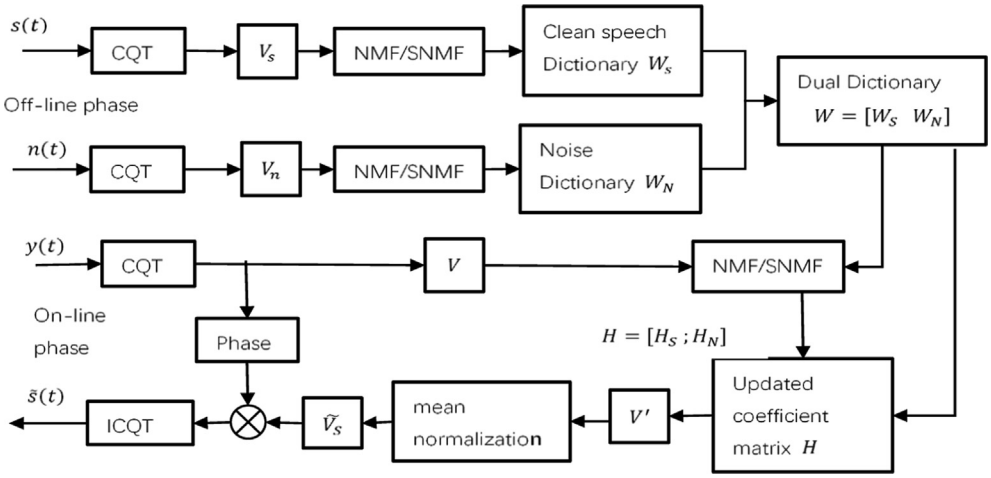

Fig. 1. The diagram of the proposed speech enhancement system based on NMF/SNMF-CQT approach.

where the product and division are carried out in an element-wise fashion. Multiplicative update rules for the activation matrices of NMF algorithm are derived, as equation (13). Multiplicative update rules for the activation matrices of SNMF algorithm are derived, as equation (14).

Then, we multiply the updated matrix H_s with the prior matrix W_s to obtain a reconstructed speech amplitude matrix V_s. The phase of the enhanced speech is assumed to be the same as the input noisy utterance. With the speech amplitude matrix and phase, we use inverse Constant Q Transform (ICQT) to produce the enhanced time-domain signal s(t). It shows our proposed speech enhancement system based on NMF or SNMF in the Constant-Q frequency domain process in Fig. 1.

5. Simulation result and conclusion

5.1. Experimental setup

The experiments were done using the LibriTTS corpus [27]. The corpus contains approximately 585 h of reading English speech at a 24 kHz sampling rate. Our experiments are gender-dependent. In the training phase, 120 utterances for female and male are randomly chosen for dictionary learning, respectively. In addition, the dictionaries W_s and W_n are obtained with the NMF. In our experiments, seven types of noise: Engine Room Noise (Destroyer), Cockpit noise 3 (F16), Factory floor noise1 (F1), Factory floor noise 2 (F2), Military Vehicle Noise (M109), Vehicle Interior Noise (Volvo), and White noise are considered [28]. Destroyer noise is acquired by recording samples from microphone onto digital audio tape (DAT), the sound level during the recording process was 101 dBA. F16 noise, factory floor noise1, factory floor noise 2, M109 noise, and Volvo noise are acquired by recording samples from 1/2″ B&K Condenser microphone onto digital audio tape (DAT). Factory floor noise 1 was recorded near plate-cutting and electrical welding equipment, factory floor noise 2 was recorded in a car production hall, M109 noise records the sound of M109 tanks was moving at a speed of 30 km/h. The sound level during the recording process was 100 dBA. Recording of Volvo was made at 120 km/h, in 4th gear, on an asphalt road, in rainy conditions. White noise was acquired by sampling a high-quality analog noise generator, which results in equal energy per Hz bandwidth. Besides, six different Signal to Noise Ratio (SNR) are set, which are −5dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. All the noises are artificially added to the clean utterance.

NMF and SNMF are used to perform noise matching and noise mismatch experiments respectively. In the noise matching experiment, the noise that corrodes clean speech has been trained in the noise dictionary during the training phase. In this process, we trained Destroyer, F16, Factory1, Factory2, M109, Volvo, and White noise. Seven kinds of noise produce seven separate noise dictionaries. In the test phase, the corresponding noise dictionary and the clean speech dictionary are used to denoise. To further illustrate speech enhancement performance based on NMF and SNMF in the CQT domain, we set up a noise mismatch experiment. In the noise mismatch experiment, the noise that corrodes clean speech is not trained in the training stage. We train Destroyer, F16, Factory2, and Volvo noises in advance to constitute a noise dictionary. In the test phase, the speech we need to denoise is corrupted by Factory1 and M109 noise respectively.

5.2. Parameter setup

In NMF-CQT and SNMF-CQT experiments, we use CQT instead of STFT to extract the frequency domain parameters of the speech signal. In the CQT method, K determines the accuracy of the frequency domain speech signal. The calculation equation of K is as follows:

$$K = B \times \log_2 \frac{f_{max}}{f_{min}} \# \tag{23}$$

OCT is the number of octaves (generally an integer less than or equal to (9):

$$OTC = \log_2 \frac{f_{max}}{f_{min}} \# \tag{24}$$

$$f_{max} = f_s/2 \# \tag{24}$$

$$f_{min} = f_s/2^{10} \# \tag{26}$$

where f_s is the sampling frequency, and B (B = 48 in this experiment) is the number of frequency bins per octave. So the size of K is determined by B and f_s. Converting the time domain signal of the speech into the frequency domain signal with STFT, we use 512 fast-Fourier transform points to characterize 24 kHz speech sounds.

In the noise matching experiment based on NMF, we set the number base of the male speech dictionary to 330, the female speech dictionary to 330, and the noise dictionary to 150, the number of iterations in the NMF algorithm is 25. In the noise mismatching based on NMF experiment, we set the number base of the male speech dictionary to 125, the female speech dictionary to 125, and

the noise dictionary to 100, the number of iterations in the NMF algorithm is 25.

In the noise matching and noise mismatch experiment based on SNMF, we set the number R of basis vectors for speech to 1000 and we set the number R of basis vectors for noise to 750. At the training phase and testing phase, the number of iterations in the SNMF algorithm is 125. We set the sparse to 1.6.

6. Performance Evaluation

We evaluate the ability of speech enhancement by the metric PESQ and STOI. The PESQ algorithm is an effective algorithm based on the input–output mode. It needs a noisy signal and an original signal. The score is mapped to the subjective mean opinion score (MOS). PESQ score ranges from − 0.5 to 4.5. The higher the score, the better the speech quality. Similar to PESQ, STOI also needs a noisy signal and an original signal. While its score ranges from 0 to 1. The higher the score, the better the intelligibility.

6.1. Noise matching experimental performance evaluation based on NMF

We shorten seven noises: Destroyer, F16, Factory1, Factory2, M109, Volvo, and White from N1 to N7 in the experimental tables. We use the STFT based spectrogram as the baseline to compare with our proposed CQT based approach. Note that both these two approaches take the current frame and adjacent four frames as the input of the NMF. In this subsection, results are presented for female and male, seven noises conditions, six SNRs under metric PESQ, and STOI.

Table 1 shows the mail trail performance for NMF method using STFT and CQT transformation approaches when speech to noise ratio (SNR) changes from −5 to 20. Results are shown separately for female and male partitions with respect to PESQ and STOI. From Table 1, we find that for the speech that corrupted by Destroyer (N1) and F16(N2) noise, the effect of speech enhancement based on NMF-CQT is better when the SNR is −5 and 0, and the effect of speech enhancement based on NMF-STFT is better when the SNR is higher than 5. In terms of speech that is corrupted by Factory1(N3), Factory2(N4), and M109(N5) noise, the NMF-CQT approach outperforms the NMF-STFT baseline when SNR is −5,0 and 5 for PESQ, and outperforms the baseline at all SNRs for STOI. When the speech is corrupted by Volvo(N6) noise, the speech enhanced based on NMF-CQT is worse than the baseline at all SNRs for PESQ, while equivalent to the baseline at all SNRs for STOI. When it comes to the White(N7) condition, we obverse that only when SNR = -5, our proposed method works better than the baseline in terms of PSEQ. When averaging the PESQ and STOI in seven noise conditions, we notice that the proposed NMF-CQT approach outperforms the baseline when SNR is −5 and 0 for PESQ, and −5,0 and 5 for STOI.

Table 2 gives a series of results of the female trail. Speech enhancement based on NMF-CQT for Destroyer, F16, Factory1, Factory2, M109 and Volvo noise has improved compared to speech enhancement based on NMF-STFT. Speech enhancement based on NMF-CQT is comparable with NMF-STFT in White noise condition. That is, the proposed NMF-CQT approach outperforms the baseline across all noise types at all SNRs for both PESQ and STOI, except for White noise condition. For White noise condition, there is a slight decrease in terms of STOI, while PESQ has improvement across all SNRs. From the average of seven types of noise, the proposed NMF-CQT approach outperforms the baseline at all SNRs for both PESQ and STOI.

From Table 1 and Table 2, we found that speech enhancement based on NMF-CQT has a better denoise effect on female speech than male speech.

In Table 3, we compare the results under different sampling rates. The aforementioned experimental setting part mentioned that the sampling rate is 24 kHz, then it is down sampled to 16 k and 8 k, respectively. It is clear that with a higher sampling rate, the enhanced speech has better quality and intelligibility. The bold value shows the best performance entries, and we notice taking our proposed method at a 24 kHz sampling rate shows the best result.

Table 1
Male trail performance comparison of the enhanced speech based on NMF-STFT and NMF-CQT when the sampling rate is 24 kHz. There are seven noise types: Destroyer-N1, F16-N2, Factory1-N3, Factory2-N4, M109-N5, Volvo-N6, and White-N7. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from NMF-STFT to NMF-CQT based on the average of PESQ and STOI of the seven noise conditions.

SNR	Noise	N1	N2	N3	N4	N5	N6	N7	Average	N1	N2	N3	N4	N5	N6	N7	Average
	Metric	PESQ								STOI							
−5	STFT	1.37	1.09	1.01	1.55	1.63	2.59	1.71	1.57	0.62	0.54	0.49	0.66	0.69	0.93	0.69	0.66
	CQT	1.50	1.19	1.22	1.78	1.84	2.57	1.76	1.69	0.64	0.56	0.54	0.71	0.73	0.93	0.67	0.68
	Absolute Improvement	0.13	0.10	0.21	0.23	0.21	−0.02	0.05	**0.13**	0.02	0.02	0.06	0.05	0.05	0.00	−0.02	**0.02**
0	STFT	1.82	1.59	1.47	1.99	2.04	2.77	2.09	1.97	0.75	0.69	0.64	0.79	0.81	0.95	0.78	0.77
	CQT	1.87	1.63	1.64	2.12	2.15	2.70	2.05	2.02	0.75	0.70	0.68	0.82	0.83	0.95	0.75	0.78
	Absolute Improvement	0.05	0.03	0.18	0.12	0.11	−0.07	−0.04	**0.05**	0.00	0.00	0.04	0.03	0.02	0.00	−0.03	**0.01**
5	STFT	2.20	2.05	1.92	2.32	2.36	2.89	2.36	2.30	0.84	0.81	0.77	0.87	0.88	0.96	0.84	0.85
	CQT	2.16	2.00	2.00	2.36	2.37	2.80	2.25	2.28	0.83	0.80	0.78	0.88	0.89	0.96	0.81	0.85
	Absolute Improvement	−0.04	−0.05	0.07	0.04	0.01	−0.10	−0.11	**−0.03**	−0.01	−0.01	0.01	0.01	0.01	0.00	−0.03	**0.00**
10	STFT	2.46	2.38	2.27	2.55	2.59	2.97	2.55	2.54	0.89	0.88	0.85	0.90	0.92	0.97	0.89	0.90
	CQT	2.36	2.28	2.25	2.53	2.54	2.87	2.39	2.46	0.88	0.87	0.85	0.92	0.93	0.96	0.86	0.89
	Absolute Improvement	−0.10	−0.10	−0.03	−0.02	−0.05	−0.10	−0.16	**−0.08**	−0.01	−0.01	0.00	0.01	0.01	0.00	−0.03	**−0.01**
15	STFT	2.63	2.60	2.51	2.71	2.74	3.02	2.68	2.70	0.92	0.92	0.89	0.92	0.94	0.97	0.92	0.93
	CQT	2.51	2.48	2.43	2.66	2.67	2.92	2.50	2.59	0.91	0.91	0.89	0.94	0.95	0.96	0.89	0.92
	Absolute Improvement	−0.12	−0.12	−0.09	−0.05	−0.08	−0.10	−0.17	**−0.11**	−0.01	−0.01	0.00	0.01	0.01	0.00	−0.03	**−0.01**
20	STFT	2.75	2.75	2.67	2.81	2.85	3.05	2.77	2.81	0.94	0.94	0.92	0.93	0.95	0.97	0.94	0.94
	CQT	2.62	2.62	2.56	2.76	2.77	2.95	2.59	2.70	0.92	0.93	0.92	0.95	0.96	0.97	0.91	0.94
	Absolute Improvement	−0.13	−0.13	−0.11	−0.05	−0.08	−0.10	−0.17	**−0.11**	−0.01	−0.01	0.00	0.02	0.01	0.00	−0.03	**0.00**

Table 2
Female trail performance comparison of the enhanced speech based on NMF-STFT and NMF-CQT when the sampling rate is 24 kHz. There are seven noise types: Destroyer-N1, F16-N2, Factory1-N3, Factory2-N4, M109-N5, Volvo-N6, and White-N7. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from NMF-STFT to NMF-CQT based on the average of PESQ and STOI of the seven noise conditions.

SNR	Noise Metric	N1	N2	N3	N4	N5 PESQ	N6	N7	Average	N1	N2	N3	N4	N5 STOI	N6	N7	Average
−5	STFT	1.46	1.11	1.21	1.74	1.69	2.70	1.78	1.67	0.61	0.53	0.51	0.69	0.69	0.91	0.64	0.66
	CQT	1.71	1.24	1.36	2.07	1.97	2.96	1.84	1.88	0.66	0.56	0.55	0.73	0.74	0.94	0.65	0.69
	Absolute Improvement	0.25	0.13	0.14	0.34	0.27	0.26	0.05	**0.21**	0.04	0.04	0.04	0.04	0.05	0.02	0.01	**0.03**
0	STFT	1.90	1.63	1.69	2.14	2.12	2.86	2.13	2.07	0.74	0.67	0.66	0.80	0.81	0.93	0.74	0.76
	CQT	2.12	1.75	1.85	2.44	2.36	3.13	2.20	2.26	0.77	0.70	0.69	0.83	0.84	0.95	0.73	0.79
	Absolute Improvement	0.23	0.13	0.16	0.29	0.24	0.27	0.06	**0.20**	0.03	0.02	0.03	0.03	0.03	0.02	−0.00	**0.02**
5	STFT	2.24	2.05	2.11	2.45	2.44	2.98	2.40	2.38	0.83	0.79	0.77	0.87	0.87	0.94	0.81	0.84
	CQT	2.45	2.19	2.25	2.71	2.66	3.25	2.47	2.57	0.85	0.80	0.79	0.89	0.90	0.96	0.80	0.86
	Absolute Improvement	0.21	0.14	0.15	0.26	0.22	0.27	0.07	**0.19**	0.02	0.02	0.02	0.02	0.03	0.02	−0.00	**0.02**
10	STFT	2.49	2.37	2.41	2.67	2.66	3.05	2.60	2.61	0.87	0.86	0.85	0.90	0.91	0.94	0.86	0.89
	CQT	2.70	2.54	2.56	2.92	2.89	3.33	2.68	2.80	0.90	0.88	0.86	0.93	0.93	0.96	0.85	0.90
	Absolute Improvement	0.21	0.16	0.15	0.25	0.23	0.28	0.08	**0.20**	0.02	0.01	0.02	0.02	0.02	0.02	−0.00	**0.02**
15	STFT	2.67	2.61	2.63	2.82	2.82	3.09	2.74	2.77	0.90	0.90	0.89	0.92	0.93	0.95	0.89	0.91
	CQT	2.90	2.80	2.80	3.09	3.07	3.38	2.83	2.98	0.92	0.92	0.91	0.95	0.95	0.97	0.89	0.93
	Absolute Improvement	0.23	0.19	0.17	0.27	0.25	0.29	0.09	**0.21**	0.02	0.02	0.02	0.03	0.03	0.02	−0.00	**0.02**
20	STFT	2.80	2.78	2.78	2.93	2.94	3.11	2.85	2.88	0.92	0.92	0.91	0.93	0.93	0.95	0.91	0.92
	CQT	3.05	3.01	2.98	3.22	3.21	3.40	2.96	3.12	0.94	0.94	0.93	0.96	0.96	0.97	0.91	0.95
	Absolute Improvement	0.25	0.23	0.20	0.29	0.27	0.30	0.11	**0.24**	0.02	0.02	0.02	0.03	0.03	0.02	0.00	**0.02**

Table 3
Female trail performance comparison of the enhanced speech based on NMF-STFT and NMF-CQT with different sampling rates. Original 24 kHz is resampled to 16 k and 8 k, respectively. The SNR is set as −5. The bold value represents the highest average value of PESQ and STOI of the seven noise conditions.

Sampling rate	Noise Metric	N1 PESQ	N2	N3	N4	N5	N6	N7	Average	N1 STOI	N2	N3	N4	N5	N6	N7	Average
24 kHz	STFT	1.46	1.11	1.21	1.74	1.69	2.70	1.78	1.67	0.61	0.53	0.51	0.69	0.69	0.91	0.64	0.66
	CQT	1.71	1.24	1.36	2.07	1.97	2.96	1.84	**1.88**	0.66	0.56	0.55	0.73	0.74	0.94	0.65	**0.69**
16 kHz	STFT	1.45	1.11	1.19	1.72	1.68	2.69	1.69	1.65	0.61	0.53	0.51	0.69	0.69	0.91	0.62	0.65
	CQT	1.74	1.24	1.32	2.09	1.96	2.92	1.75	1.86	0.66	0.57	0.54	0.73	0.74	0.93	0.63	0.69
8 kHz	STFT	1.45	1.08	1.15	1.72	1.66	2.69	1.38	1.59	0.61	0.52	0.5	0.69	0.68	0.91	0.55	0.64
	CQT	1.76	1.28	1.27	2.11	1.97	2.82	1.62	1.83	0.66	0.57	0.53	0.72	0.73	0.92	0.58	0.67

Overall, the speech enhancement based on NMF-CQT has a greater improvement than baseline at 24 kHz. Performance based on female data gives more support on NMF-CQT approach than the baseline.

6.2. Noise mismatching experimental performance evaluation based on NMF

Table 4 compares our proposed method and baseline under noise mismatching condition. When looking at the male part in Table 4, we find that the NMF-CQT approach outperforms the baseline for Factory 1 and M109 when SNR is −5, 0, and 5 in terms of PESQ and outperforms the baseline across all SNRs in terms of STOI. The average of these two types of noise has the same trend with the single noise, that is, the proposed NMF-CQT approach outperforms the baseline when SNR is −5,0 and 5 for PESQ, and at all SNRs for STOI.

When it comes to the female part in Table 4, we observe that the NMF-CQT approach outperforms the baseline for Factory 1 and M109 noise conditions when SNR is −5, and 0 for PESQ, and gives comparable results with baseline at all SNRs for STOI. When it comes to the average of these two types of noise, the proposed NMF-CQT approach outperforms the baseline when SNR is −5 and 0 for PESQ, and all SNRs for STOI.

From Table 4, we find that in the case of noise mismatching, NMF-CQT approach is still better than NMF-STFT at low SNR. Comparing Table 4 with Tables 1 and 2, it is obvious that the denoising effect of noise mismatch is worse than that of noise matching condition. Therefore, the effect of supervised denoising is better than that of unsupervised denoising. In the future denoising process, due to the complex environment, we should increase the number of noise dictionaries as much as possible, or adopt adaptive methods to avoid unsupervised conditions.

6.3. Noise matching experimental performance evaluation based on SNMF

Table 5 shows the mail trail performance for NMF method using STFT-CQT approach when speech to noise ratio (SNR) changes from −5 to 20. Results are shown separately for female and male partitions with respect to PESQ and STOI. From Table 5, we find that for the speech that corrupted by Destroyer(N1) noise, the SNMF-CQT approach outperforms the SNMF-STFT approach when SNR is −5,0 and 5 for PESQ, and outperforms the SNMF-STFT when the SNR is −5 for STOI, the SNMF-CQT approach is almost equal with SNMF-STFT approach when SNR is higher than – 5 for STOI. For the speech that is corrupted by F16(N2) noise, the effect of speech enhancement based on SNMF-CQT is better when the SNR is −5, and the effect of speech enhancement based on SNMF-STFT is better when the SNR is higher than 0. For the speech that corrupted by Factory1(N3) noise, the speech enhanced based on SNMF-CQT is worse than the baseline at all SNRs for both PESQ and STOI. In terms of speech that is corrupted by Factory2(N4), M109(N5) and Volvo(N6) noise, the SNMF-CQT approach outperforms the SNMF-STFT approach at all SNRs for both PESQ and STOI. When it comes to the White(N7) condition, we obverse that only when SNR = -5, our proposed method works better than the baseline in terms of PESQ. When averaging the PESQ and STOI in seven noise

Table 4
Performance comparison of the enhanced speech based on NMF-STFT and NMF-CQT when the sampling rate is 24 kHz. Two noise types: Factory1-N1, M109-N5 are taken as the evaluation condition. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from NMF-STFT to NMF-CQT based on the average of PESQ and STOI of the two noise conditions. Left part shows male trail performance and right gives female trail performance.

SNR	Gender	Male						Female					
	Noise	N1	N2	Average	N1	N2	Average	N1	N2	Average	N1	N2	Average
	Metric	PESQ			STOI			PESQ			STOI		
−5	STFT	0.86	1.58	1.22	0.44	0.59	0.52	0.89	1.51	1.20	0.51	0.66	0.58
	CQT	0.97	1.72	1.34	0.49	0.64	0.57	0.94	1.54	1.24	0.52	0.67	0.60
	Absolute Improvement	0.11	0.14	**0.12**	0.05	0.05	**0.05**	0.04	0.04	**0.04**	0.02	0.01	**0.02**
0	STFT	1.13	1.87	1.50	0.59	0.73	0.66	1.36	1.93	1.64	0.65	0.78	0.71
	CQT	1.23	1.96	1.60	0.63	0.76	0.69	1.37	1.93	1.65	0.65	0.79	0.72
	Absolute Improvement	0.10	0.09	**0.10**	0.04	0.03	**0.04**	0.01	0.00	**0.01**	0.01	0.00	**0.01**
5	STFT	1.57	2.15	1.86	0.72	0.82	0.77	1.82	2.28	2.05	0.77	0.86	0.82
	CQT	1.63	2.18	1.90	0.74	0.84	0.79	1.82	2.27	2.04	0.77	0.86	0.81
	Absolute Improvement	0.06	0.03	**0.04**	0.02	0.02	**0.02**	−0.01	−0.01	**−0.01**	0.00	0.00	**0.00**
10	STFT	1.99	2.37	2.18	0.81	0.87	0.84	2.22	2.56	2.39	0.86	0.90	0.88
	CQT	1.98	2.35	2.17	0.82	0.88	0.85	2.21	2.55	2.38	0.85	0.90	0.88
	Absolute Improvement	0.00	−0.02	**−0.01**	0.01	0.01	**0.01**	−0.01	−0.02	**−0.01**	0.00	0.00	**0.00**
15	STFT	2.29	2.53	2.41	0.87	0.90	0.89	2.52	2.78	2.65	0.90	0.92	0.91
	CQT	2.25	2.49	2.37	0.88	0.91	0.89	2.51	2.76	2.64	0.90	0.92	0.91
	Absolute Improvement	−0.04	−0.05	**−0.04**	0.01	0.01	**0.01**	0.00	0.00	**0.00**	0.00	0.00	**0.00**
20	STFT	2.50	2.65	2.58	0.90	0.91	0.91	0.93	0.94	0.93	0.93	0.94	0.93
	CQT	2.44	2.60	2.52	0.91	0.92	0.91	0.92	0.93	0.93	0.92	0.93	0.93
	Absolute Improvement	−0.06	−0.05	**−0.06**	0.01	0.00	**0.01**	0.00	0.00	**0.00**	0.00	0.00	**0.00**
Improvement	−0.06	−0.05	**−0.06**	0.01	0.00	**0.01**	0.00	0.00	**0.00**	0.00	0.00	**0.00**	

Table 5
Male trail performance comparison of the enhanced speech based on SNMF-STFT and SNMF-CQT when the sampling rate is 24 kHz. There are seven noise types: Destroyer-N1, F16-N2, Factory1-N3, Factory2-N4, M109-N5, Volvo-N6, and White-N7. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from SNMF-STFT to SNMF-CQT based on the average of PESQ and STOI of the seven noise conditions.

SNR	Noise	N1	N2	N3	N4	N5	N6	N7	Average	N1	N2	N3	N4	N5	N6	N7	Average
	Metric	PESQ								STOI							
−5	STFT	1.54	1.09	1.14	1.61	1.69	2.60	1.49	1.59	0.70	0.59	0.56	0.71	0.74	0.92	0.71	0.70
	CQT	1.76	1.14	1.03	1.94	2.05	2.64	1.51	1.72	0.72	0.59	0.55	0.76	0.80	0.94	0.68	0.72
	Absolute Improvement	0.22	0.05	−0.11	0.33	0.35	0.04	0.02	**0.13**	0.02	0.00	−0.01	0.05	0.06	0.02	−0.03	**0.02**
0	STFT	1.93	1.51	1.65	2.05	2.07	2.73	1.89	1.98	0.81	0.73	0.71	0.82	0.83	0.94	0.79	0.80
	CQT	2.06	1.46	1.49	2.23	2.29	2.77	1.84	2.02	0.81	0.72	0.69	0.84	0.87	0.96	0.76	0.81
	Absolute Improvement	0.14	−0.05	−0.16	0.18	0.22	0.03	−0.05	**0.04**	0.00	−0.01	−0.02	0.03	0.04	0.01	−0.03	**0.00**
5	STFT	2.24	1.92	2.06	2.35	2.36	2.84	2.22	2.29	0.87	0.83	0.81	0.88	0.89	0.96	0.86	0.87
	CQT	2.29	1.81	1.89	2.43	2.46	2.87	2.12	2.27	0.87	0.82	0.80	0.90	0.91	0.97	0.83	0.87
	Absolute Improvement	0.05	−0.11	−0.18	0.08	0.10	0.03	−0.10	**−0.02**	0.00	−0.01	−0.01	0.02	0.02	0.01	−0.03	**0.00**
10	STFT	2.47	2.25	2.36	2.56	2.56	2.91	2.47	2.51	0.91	0.89	0.88	0.92	0.92	0.96	0.90	0.91
	CQT	2.46	2.11	2.20	2.59	2.61	2.94	2.34	2.46	0.91	0.89	0.88	0.93	0.94	0.97	0.88	0.91
	Absolute Improvement	−0.01	−0.13	−0.16	0.03	0.05	0.03	−0.13	−0.05	0.00	−0.01	0.00	0.01	0.02	0.01	−0.02	**0.00**
15	STFT	2.62	2.49	2.56	2.69	2.70	2.96	2.64	2.67	0.94	0.93	0.91	0.94	0.94	0.97	0.93	0.94
	CQT	2.59	2.36	2.44	2.71	2.73	2.99	2.52	2.62	0.94	0.93	0.92	0.95	0.96	0.97	0.92	0.94
	Absolute Improvement	−0.03	−0.13	−0.11	0.02	0.03	0.03	−0.13	**−0.05**	0.00	0.00	0.01	0.01	0.01	0.01	−0.02	**0.00**
20	STFT	2.74	2.67	2.69	2.79	2.80	2.99	2.76	2.78	0.95	0.95	0.93	0.95	0.95	0.97	0.95	0.95
	CQT	2.71	2.56	2.62	2.82	2.83	3.02	2.66	2.75	0.95	0.95	0.95	0.96	0.97	0.97	0.94	0.96
	Absolute Improvement	−0.03	−0.11	−0.07	0.02	0.03	0.03	−0.10	**−0.03**	0.00	0.00	0.01	0.01	0.01	0.00	−0.01	**0.00**

conditions, we notice that the SNMF-CQT approach outperforms the SNMF-STFT when SNR is −5 and 0 for both PESQ and STOI.

Table 6 gives a series of results of the female trail. Speech enhancement based on SNMF-CQT under Destroyer, F16, Factory2, M109 and Volvo noise has improved compared to speech enhancement based on SNMF-STFT. When the speech is corrupted by Factory1(N3) noise, the speech enhanced based on SNMF-CQT approach better than SNMF-STFT approach when SNR is 10,15 and 20 for both PESQ and STOI, but the speech enhanced based on SNMF-CQT approach worse than SNMF-STFT approach when SNR is −5,0 and 10 for both PESQ and STOI. When it comes to the White(N7) condition, we obverse that our proposed method works better than the baseline at all SNRs in terms of PESQ, and outperforms the baseline when the SNR is 15 and 20 for STOI. From the average of seven types of noise, the proposed SNMF-CQT approach outperforms the SNMF-STFT approach at all SNRs for both PESQ and STOI.

From Table 5 and Table 6, we found that speech enhancement based on SNMF-CQT has a better denoise effect on female speech than male speech.

6.4. Noise mismatching experimental performance evaluation based on SNMF

Table 7 compares our proposed method and baseline under noise mismatching condition. When looking at the male part in Table 7, we find that the male speech enhancement based on SNMF-CQT approach outperforms the baseline for Factory 1 and M109 when SNR is −5, 0, and 5 in terms of PESQ and outperforms the baseline across all SNRs in terms of STOI. The average of these two types of noise has the same trend with the single noise, that is, the proposed NMF-CQT approach outperforms the baseline when SNR is −5,0 and 5 for PESQ, and at all SNRs for STOI. From the results of the female trail, the proposed SNMF-CQT approach out-

Table 6
Female trail performance comparison of the enhanced speech based on SNMF-STFT and SNMF-CQT when the sampling rate is 24 kHz. There are seven noise types: Destroyer-N1, F16-N2, Factory1-N3, Factory2-N4, M109-N5, Volvo-N6, and White-N7. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from SNMF-STFT to SNMF-CQT based on the average of PESQ and STOI of the seven noise conditions.

SNR	Noise Metric	N1 PESQ	N2	N3	N4	N5	N6	N7	Average	N1 STOI	N2	N3	N4	N5	N6	N7	Average
−5	STFT	1.56	1.04	1.34	1.82	1.78	2.65	1.49	1.67	0.67	0.56	0.57	0.71	0.72	0.90	0.66	0.68
	CQT	2.01	1.09	1.14	2.22	2.32	3.02	1.59	1.91	0.73	0.57	0.56	0.77	0.80	0.94	0.66	0.72
	Absolute Improvement	0.46	0.05	−0.20	0.40	0.54	0.37	0.11	**0.24**	0.05	0.01	−0.01	0.06	0.08	0.04	0.00	**0.03**
0	STFT	1.94	1.50	1.81	2.17	2.12	2.80	1.88	2.03	0.78	0.69	0.71	0.81	0.82	0.92	0.75	0.78
	CQT	2.32	1.55	1.68	2.53	2.58	3.18	1.94	2.25	0.81	0.70	0.69	0.85	0.87	0.95	0.74	0.80
	Absolute Improvement	0.38	0.04	−0.12	0.36	0.46	0.38	0.06	**0.22**	0.03	0.00	−0.02	0.04	0.05	0.04	−0.01	**0.02**
5	STFT	2.23	1.91	2.14	2.42	2.37	2.90	2.19	2.31	0.85	0.80	0.80	0.87	0.87	0.93	0.82	0.85
	CQT	2.56	1.98	2.13	2.76	2.80	3.30	2.23	2.54	0.87	0.80	0.80	0.90	0.91	0.96	0.81	0.87
	Absolute Improvement	0.33	0.06	−0.01	0.35	0.42	0.40	0.03	**0.23**	0.03	0.01	−0.01	0.03	0.04	0.03	−0.01	**0.02**
10	STFT	2.45	2.22	2.39	2.60	2.57	2.97	2.44	2.52	0.89	0.86	0.86	0.90	0.90	0.94	0.87	0.89
	CQT	2.76	2.34	2.48	2.96	2.98	3.39	2.48	2.77	0.91	0.88	0.87	0.93	0.94	0.97	0.87	0.91
	Absolute Improvement	0.31	0.11	0.09	0.36	0.40	0.42	0.04	0.25	0.03	0.01	0.01	0.03	0.03	0.03	0.00	**0.02**
15	STFT	2.61	2.47	2.58	2.73	2.72	3.01	2.63	2.68	0.91	0.90	0.90	0.92	0.92	0.94	0.90	0.91
	CQT	2.93	2.63	2.76	3.12	3.13	3.44	2.70	2.96	0.94	0.93	0.92	0.95	0.96	0.97	0.91	0.94
	Absolute Improvement	0.32	0.17	0.18	0.39	0.40	0.43	0.08	**0.28**	0.03	0.02	0.02	0.03	0.03	0.03	0.01	**0.03**
20	STFT	2.74	2.64	2.71	2.84	2.83	3.04	2.76	2.79	0.92	0.92	0.92	0.93	0.93	0.94	0.92	0.93
	CQT	3.08	2.88	2.99	3.26	3.25	3.46	2.90	3.12	0.95	0.95	0.95	0.96	0.96	0.97	0.94	0.95
	Absolute Improvement	0.35	0.24	0.28	0.42	0.42	0.42	0.14	**0.32**	0.03	0.03	0.03	0.03	0.03	0.03	0.02	**0.03**

Table 7
Performance comparison of the enhanced speech based on SNMF-STFT and SNMF-CQT when the sampling rate is 24 kHz. Two noise types: Factory1-N1, M109-N5 are taken as the evaluation condition. The SNR is set from −5 to 20 at interval 5. The bold value represents the absolute improvement from NMF-STFT to SNMF-CQT based on the average of PESQ and STOI of the two noise conditions. Left part shows male trail performance and right gives female trail performance.

SNR	Gender Noise Metric	Male N1 PESQ	N2	Average	N1 STOI	N2	Average	Female N1 PESQ	N2	Average	N1 STOI	N2	Average
−5	STFT	1.43	1.14	1.28	0.67	0.57	0.62	1.47	1.13	1.3	0.66	0.55	0.61
	CQT	1.63	1.34	1.49	0.7	0.61	0.66	1.89	1.46	1.68	0.7	0.61	0.66
	Absolute Improvement	0.21	0.19	**0.2**	0.03	0.04	**0.03**	0.42	0.34	**0.38**	0.05	0.06	**0.05**
0	STFT	1.83	1.6	1.72	0.78	0.72	0.75	1.85	1.62	1.74	0.77	0.7	0.73
	CQT	1.96	1.74	1.85	0.8	0.73	0.76	2.23	1.92	2.07	0.8	0.73	0.76
	Absolute Improvement	0.13	0.14	**0.13**	0.01	0.02	**0.01**	0.37	0.3	**0.34**	0.03	0.04	**0.03**
5	STFT	2.17	2.01	2.09	0.86	0.82	0.84	2.16	2	2.08	0.84	0.79	0.82
	CQT	2.22	2.06	2.14	0.86	0.82	0.84	2.49	2.27	2.38	0.86	0.82	0.84
	Absolute Improvement	0.04	0.05	**0.05**	0.00	0.00	**0.00**	0.33	0.28	**0.31**	0.03	0.03	**0.03**
10	STFT	2.42	2.31	2.36	0.9	0.88	0.89	2.38	2.28	2.33	0.88	0.86	0.87
	CQT	2.41	2.3	2.35	0.91	0.88	0.9	2.71	2.55	2.63	0.91	0.88	0.9
	Absolute Improvement	−0.01	−0.01	**−0.01**	0.00	0.00	**0.00**	0.33	0.28	**0.30**	0.03	0.03	**0.03**
15	STFT	2.59	2.52	2.55	0.93	0.92	0.92	2.56	2.48	2.52	0.9	0.89	0.9
	CQT	2.56	2.48	2.52	0.93	0.92	0.93	2.89	2.78	2.84	0.93	0.92	0.93
	Absolute Improvement	−0.03	−0.04	**−0.03**	0.01	0.00	**0.00**	0.34	0.29	**0.32**	0.03	0.03	**0.03**
20	STFT	2.71	2.67	2.69	0.94	0.94	0.94	2.69	2.64	2.66	0.92	0.91	0.92
	CQT	2.68	2.63	2.66	0.95	0.94	0.95	3.05	2.97	3.01	0.95	0.94	0.95
	Absolute Improvement	−0.03	−0.04	**−0.03**	0.01	0.01	**0.01**	0.36	0.33	**0.35**	0.03	0.03	**0.03**

performs the baseline at all SNRs for both PESQ and STOI. From Table 7, we find that in the case of noise mismatching, SNMF-CQT approach is still better than SNMF-STFT at low SNR. Comparing Table 7 with Table 5 and 6, it is obvious that the denoising effect of noise mismatch is worse than that of noise matching condition.

6.5. Comparison between NMF and SNMF based methods

Comparing table 1 and 2 with table 5 and table 6, we find that in the noise matching experiment, for the speech that corrupted by Destroyer, Factory2 and M109 noise, the effect of speech enhancement based on SNMF algorithm outperforms the that of NMF algorithm at low SNRs for both PESQ and STOI, and from average, for the speech that corrupted by seven type noise , the effect of speech enhancement based on SNMF algorithm outperforms the that of NMF algorithm at low SNRs for PESQ and at all SNRs for STOI. Comparing table 4 with table 7, we find that in the noise mismatching experiment, when the speech is corrupted by Factory1 noise, the speech enhanced based on SNMF algorithm is better than the NMF algorithm at all SNRs for both PESQ and STOI, when the speech is corrupted by M109 noise, the speech enhanced based on SNMF algorithm is worse than the NMF algorithm at all SNRs for PESQ, from the average of two types of noise, the speech enhanced based on SNMF algorithm outperforms NMF algorithm at all SNRs for both PESQ and STOI. Moreover, compared with NMF algorithm, SNMF algorithm is more effective in frequency domain sampling using CQT.

From the average of all results, for the male trail, whether it is noise matching or noise mismatch, the denoise effect of NMF-CQT approach is better than that of NMF-STFT approach at low SNR, and the denoise effect of SNMF-CQT approach is better than that of SNMF-STFT approach at low SNR. For the female trail, whether it is noise matching or noise mismatch, the denoise effect

of NMF-CQT approach is better than that of NMF-STFT approach at all SNRs, and the denoise effect of SNMF-CQT approach is better than that of SNMF-STFT approach at all SNRs.

7. Conclusion

In this paper, the NMF algorithm and the SNMF algorithm are used respectively. From the experimental results, the denoising effects of the two approaches are different for different noises. From the average of results, when we denoise male speech corrupted by noise, using CQT is better than STFT under the condition of low SNRs. When we denoise female speech corrupted by noise, using CQT is better than STFT at all SNRs. Comparing with the NMF algorithm, the denoising effect of SNMF algorithm is better at low SNRs for PESQ and at all SNRs for STOI. In a word, whether NMF algorithm or SNMF algorithm, using the CQT to extract the frequency domain spectrum of the speech signal is better than that of STFT in the process of speech denoising, especially in low SNR. The advantage of CQT is that it increases the low-frequency resolution of the speech, and the human utterance is concentrated in the low frequency.

CRediT authorship contribution statement

Longting Xu: Conceptualization, Methodology, Software. **Zhilin Wei:** Data curation, Writing - original draft preparation. **Syed Faham Ali Zaidi:** Writing - review & editing, Investigation. **Bo Ren:** Supervision, Software, Validation. **Jichen Yang:** Writing-Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Scalart P, Filho J. Speech enhancement based on a priori signal to noise estimation. Proc IEEE Int Conf Acoust Speech, Sign Process 1996:629–32.
[2] Cohen I. Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator. IEEE Signal Process Lett 2002;9(4):113–6.
[3] Loizou PC. Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. IEEE Trans Audio Speech Lang Process 2005;13(5):857–69.
[4] Chen Y, Shi L, Feng Q, Yang J, Shu H, Luo L, Coatrieux J-L, Chen W. Artifact Suppressed Dictionary Learning for Low-Dose CT Image Processing. IEEE Trans Med Imaging 2014;33(12):2271–92. https://doi.org/10.1109/TMI.2014.2336860.
[5] Chen Y, Zhang Y, Shu H, Yang J, Luo L, Coatrieux J-L, Feng Q. Structure-Adaptive Fuzzy Estimation for Random-Valued Impulse Noise Suppression. IEEE Trans Circuits Syst Video Technol 2018;28(2):414–27. https://doi.org/10.1109/TCSVT.2016.2615444.
[6] Aharon M, Elad M, Bruckstein A. $rm K$—SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. IEEE Trans Signal Process 2006;54(11):4311–22.
[7] Sigg CD, Dikk T, Buhmann JM. Speech Enhancement Using Generative Dictionary Learning. IEEE Trans Audio Speech Lang Process 2012;20(6):1698–712. https://doi.org/10.1109/TASL.2012.2187194.
[8] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401(6755):788–91.
[9] Mohammadiha N, Smaragdis P, Leijon A. Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization. IEEE Trans Audio Speech Lang Process 2013;21(10):2140–51.
[10] Févotte C, Bertin N, Durrieu J-L. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. Neural Comput 2009;21(3):793–830. https://doi.org/10.1162/neco.2008.04-08-771.
[11] Kwon K, Jong WS, Nam SK. NMF-based speech enhancement using bases update. IEEE Sig Process Lett 2015;22(4):450–4.
[12] Zhou WL et al. Speech denoising using Bayesian NMF with online base update. Multimed Tool Appl 2019;78(11):15647–64.
[13] Qing Z, Zuren F, Emmanouil B. Adaptive Noise Reduction for Sound Event Detection Using Subband-Weighted NMF. Sensors (Basel, Switzerland) 2019;19(14).
[14] Todisco, M., et al. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. Speaker and Language Recognition Workshop, Odyssey 2016, June 21, 2016 - June 24, 2016, Bilbao, Spain, International Speech Communication Association; 2016.
[15] Huang, S., et al. Timbretron: A wavenet(Cyclegan(CqT(Audio))) pipeline for musical timbre transfer. 7th International Conference on Learning Representations, ICLR 2019, May 6, 2019 - May 9, 2019, New Orleans, LA, United states, International Conference on Learning Representations, ICLR; 2019.
[16] Diniz, F. C. C. B., et al. "High-selectivity filter banks for spectral analysis of music signals." Eurasip J Adv Signal Process; 2007.
[17] Yang JC et al. Extraction of Octave Spectra Information for Spoofing Attack Detection. Ieee-Acm Trans Audio Speech Language Process 2019;27(12):2373–84.
[18] Yang JC, Das RK. Improving anti-spoofing with octave spectrum and short-term spectral statistics information. Appl Acoust 2020;157:107017.
[19] Fu S-W et al. Joint Dictionary Learning-Based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility after Oral Surgery. IEEE Trans Biomed Eng 2017;64(11):2584–94.
[20] Jonathan Le Roux, Felix Weninger, John R. Hershey, "Sparse NMF – half-baked or well done?," Mitsubishi Electric Research Laboratories Technical Report, TR2015-023, Mar; 2015.
[21] Zhou WL, He QH, Wang YL, et al. Sparse representation-based quasi-clean speech construction for speech quality assessment under complex environments. IET Signal Process 2017;11(4):486–93.
[22] Févotte.C.; Vincent, E.; Ozerov,A. Single-channel audio source separation with NMF: Divergences,constraints and algorithms. In Audio Source Separation; Makino, S., Ed.; Springer: Cham, Switzerland, 2018; pp. 1–24.
[23] Brown JC. Calculation of a constant Q spectral transform. J Acoust Soc Am 1991;89(1):425–34.
[24] Dobre, R.A., Negrescu, C. Automatic music transcription software based on constant Q transform. Proc. 8th Int. Conf. Electron. Comput Artif Intell ECAI 2016; 2017. Doi: 10.110 9/ECAI.2016.7861193.
[25] Zhizhong Ding. Fast Algorithm of CQ Transformation and Error Analysis in Pitch Frequency Estimation. Inform Electron Eng 2005;3(04):33–7.
[26] Brown JC, Puckette MS. An efficient algorithm for the calculation of a constant Q transform. J Acoust Soc Am 1992;92(5):2698–701.
[27] http://spib.linse.ufsc.br/database.html.
[28] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv; 2019.
[29] Amari S-I. Natural gradient works efficiently in learning. Neural Comput 1998;10(2):251–76.