Single-Channel Speech Enhancement Based on Non-negative Matrix Factorization and Online Noise Adaptation

Kwang Myung Jeon\(^1\), Chan Jun Chun\(^1\), Woo Kyeong Seong\(^1\), Hong Kook Kim\(^1\), and Myung Kyu Choi\(^2\)

\(^1\)Gwangju Institute of Science and Technology (GIST), Korea
\(^2\)Samsung Electronics, Korea

\{kmjeon, cjunhun, wkseong, hongkookj\}@gist.ac.kr, core.choi@samsung.com

Abstract

In this paper, we demonstrate a simulator for real-time speech enhancement based on a non-negative matrix factorization (NMF) technique. In particular, we propose an online noise adaptation method in an NMF framework, which is activated during non-speech intervals and used for adapting noise bases for NMF. Thus, incoming noisy speech is decomposed by using such adapted noise bases and universal speech bases that can be developed through training with examples of speech data. It is shown from the experiments that the proposed method improves speech separation performance and perceptual speech quality.

Index Terms: speech enhancement, non-negative matrix factorization, online noise adaptation, MMSE filtering

1. Introduction

Obtaining a clean speech signal in noisy environments is highly demanded for various applications, including mobile communications, voice services on smart TVs, car speech interfaces, etc. Most consumer applications use push-to-talk (PTT) interfaces to avoid malfunctions due to noise interference [1]. However, speech enhancement is still demanded, even with PTT interfaces, since noise interference remains in speech intervals. Various efforts have been made to enhance speech signals contaminated by environmental noise, where the noise spectrum was estimated using an a priori signal [2], minimum statistics (MS) [3], and a minima-controlled recursive algorithm [4]. Those methods work successfully when noise is assumed to be stationary. However, they might fail to reduce real-world noise composed of both stationary and non-stationary noise. As an alternative, non-negative matrix factorization (NMF)-based speech enhancement methods have been proposed to reduce non-stationary noise [5]. However, the performance of NMF-based speech enhancement methods degrades substantially when there is a mismatch in noise type between noise basis training and speech enhancement [5]. Such a limitation makes NMF-based speech enhancement methods less practical under real-world noise conditions. Hence, another strategy is needed to enhance speech corrupted by real-world noise.

In this paper, a single-channel NMF-based speech enhancement method is proposed to enhance speech under both stationary and non-stationary noise conditions by adapting noise bases on the fly. In the proposed method, the microphone is always on, while speech only comes through when PTT is on. Thus, noise adaptation is performed when PTT is off, i.e., during a non-speech interval. On the other hand, NMF decomposition from a noisy speech spectrum into individual speech and noise magnitude spectrum is performed when PTT is on. Next, the separated magnitude spectra are used to construct a minimum mean squared error (MMSE) filter [6]. Finally, the MMSE filter is applied to the noisy speech to estimate the clean speech.

2. Proposed speech enhancement method

Fig. 1 shows the procedure of the proposed speech enhancement method. First, the input signal at the \(i\)-th frame, \(y_i(n)\), is assumed to be \(y_i(n) = s_i(n) + d_i(n)\), where \(s_i(n)\) and \(d_i(n)\) are clean speech and additive background noise, respectively. Then, \(y_i(n)\) is transformed into \(K\) spectral components, \(Y_i(k)\), by applying a short-time Fourier transform (STFT). Next, the spectral magnitudes of \(L\) input frames are concatenated to construct a \(K \times L\) time-frequency (TF) matrix, \(Y = [Y_{i1}(k) \cdots Y_{iK}(k)]\). Depending on whether the current frame belongs to a non-speech (PTT off) or speech (PTT on) interval, \(Y_i\) is used for either noise adaptation or speech enhancement, respectively.

2.1. Online noise adaptation

When the PTT interface declares the current state to be a non-speech interval, \(Y_i\) is considered as \(i\)-th TF block of environment noise, \(D_i\). \(D_i\) is then used to update the noise basis, \(B_{b,j}\), where \(B_{b,j}\) is a \(K \times r\) matrix composed of \(K \times r\) noise atom matrices, \(b_{b,j}\), such as \(B_{b,j} = [b_{b,j,1} \cdots b_{b,j,r}]\). Then, \(b_{b,j}\) is updated through a multiplicative update rule [5], as

\[
b_{b,j} = b_{b,j} \odot \frac{D_{b,j}}{b_{b,j}^T a_{b,j}} \left( a_{b,j}^T \right)^{-1},
\]

\[
a_{b,j} = a_{b,j} \odot \frac{D_{b,j}}{b_{b,j}^T b_{b,j}} \left( b_{b,j} \right)^{-1} I\]

where \(h\) is an iteration index and both multiplication, \(\odot\), and division are applied on an element-by-element basis. In addition, \(a_{b,j}\) is an \(r \times L\) activation matrix for \(b_{b,j}\), and \(I\) is a \(K \times L\) matrix in which all elements are equal to unity. Note that all the elements of \(b_{b,j}\) and \(a_{b,j}\) are set initially at...
random values between 0 and 1. In the noise basis adaptation, \( b_{\text{noise}} \) is obtained by repeating (1) and (2) until the relative reduction of the Kullback-Leibler (KL) divergence [5] according to the iteration arrives at a value below a predefined threshold.

2.2. NMF-based speech enhancement

As shown in Fig. 1, \( Y_i \) is considered as the \( i \)-th TF block for noisy speech, \( Y_i = S_i + D_i \), and the NMF decomposition followed by MMSE filtering is performed using the updated noise bases, \( B_{\text{noise}} \), and the universal speech bases, \( B_s \). In this paper, \( B_{\text{noise}} \) is preliminarily obtained by off-line NMF training [7]. Next, clean speech and noise spectra at the \( i \)-th frame are estimated using \( \hat{S}_i = B_s \hat{a}_{\text{speech}} \) and \( \hat{D}_i = B_{\text{noise}} \hat{a}_{\text{noise}} \), respectively. The activation matrices, \( a_{\text{speech}} \) and \( a_{\text{noise}} \), are iteratively obtained using the equation of

\[
\begin{align*}
\hat{a}_{\text{speech}}^{i+1} &= \hat{a}_{\text{speech}}^{i} \odot \frac{B_s}{B_s \hat{a}_{\text{speech}}^{i} + \alpha}|Y_i|^{\gamma} \frac{1}{B_s B_s^T} \\
\hat{a}_{\text{noise}}^{i+1} &= \hat{a}_{\text{noise}}^{i} \odot \frac{B_{\text{noise}}}{B_{\text{noise}} \hat{a}_{\text{noise}}^{i} + \alpha}|Y_i|^\gamma \frac{1}{B_{\text{noise}} B_{\text{noise}}^T}
\end{align*}
\]  

(3)

where the iteration, \( h \), is also terminated when the relative reduction of the KL divergence is below a predefined threshold. Note here that the dimensions of \( B_s \), \( a_{\text{speech}} \), and \( a_{\text{noise}} \) are \( K \times R \), \( R \times L \), and \( rT \times L \), respectively, and all elements for \( a_{\text{speech}} \) and \( a_{\text{noise}} \) are randomly set between 0 and 1.

Next, an MMSE filter, \( H_i \), is constructed as a function of a smoothed signal-to-noise ratio (SNR) gain using the decision directed approach such as

\[
H_i = \frac{\xi_i}{\xi_i + 1}
\]

(4)

where \( \xi_i = \gamma \hat{S}_i E(\hat{D}_i) + (1 - \gamma) \hat{S}_i E(D_i) \) and \( \hat{S}_i \) corresponds to the estimate of the clean speech magnitude spectrum at the \( (i-1) \)-th TF frame. Thus, \( H_i \) is applied to the input spectrum, \( Y_i \), to estimate the clean speech spectrum at the \( i \)-th frame, as \( \tilde{S}_i = H_i \odot Y_i \), and finally the clean speech, \( \tilde{S}(n) \), is estimated by applying an inverse STFT to \( \tilde{S}_i \).

3. Performance evaluation

The performance of the proposed method was evaluated by measuring the source-to-interferences ratio (SIR) [8] and the mean opinion score (MOS) of the perceptual quality, and it was compared with those of conventional speech enhancement methods based on MS [3] and NMF without noise basis adaptation (NMF-Conv) [7]. The parameters of the proposed method, \( K \), \( L \), \( r \), \( T \), \( R \), and \( \gamma \) were set to 1024, 6, 4, 100, 400, and 0.6, respectively, in this paper.

For the evaluation, five males and five females spoke 10 sentences each. Each of the 100 sentences was mixed with background noise recorded in a living room with a TV on, by setting SNR at 0, 5, 10, or 15 dB. The speech and noise signals used in the evaluation were sampled at 16 kHz with a 16-bit resolution. Twenty trained listeners aged 20–30 participated in the MOS test with hi-fi headphones in a quiet room.

Table 1 compares the SIR and MOS of the proposed method with those of the conventional methods. As shown in the table, the proposed method achieved significantly higher average SIR values and average MOS than the other methods.

4. Demonstration and conclusion

We designed a simulator that employed the proposed NMF-based speech enhancement method. The simulator operated on a laptop with a clock cycle of 1.86 GHz and a memory size of 2 GB. It had a graphical user interface (GUI) for the interactive demonstration, as shown in Fig. 2. The proposed method was implemented using a Matlab code and the PTT interface control was assigned as a mouse click for the simulator. Here, we used the microphone equipped in the laptop as a speech input device. It was revealed from the real-time operation of the simulator that the proposed method improved both the speech separation performance and the perceived speech quality, compared to the conventional methods that were also implemented on the simulator.

5. Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the government of Korea (MSIP) (No. 2012-010636) and by the MSIP under the ITRC support program (NIPA-2014-H0301-14-1019) supervised by the NIPA.

6. References