Hindawi Publishing Corporation
Advances in Acoustics and Vibration
Volume 2015, Article ID 170183, 9 pages
http://dx.doi.org/10.1155/2015/170183

Research Article

Eigennoise Speech Recovery in Adverse Environments with Joint Compensation of Additive and Convolutive Noise

Trung-Nghia Phung, Huy-Khoi Do, Van-Tao Nguyen, and Quang-Vinh Thai

Thai Nguyen University of Information and Communication Technology, Thai Nguyen 250000, Vietnam

Correspondence should be addressed to Trung-Nghia Phung; ptnghia@ictu.edu.vn

Received 30 June 2015; Accepted 13 October 2015

Academic Editor: Marc Asselineau

The learning-based speech recovery approach using statistical spectral conversion has been used for some kind of distorted speech as alaryngeal speech and body-conducted speech (or bone-conducted speech). This approach attempts to recover clean speech (undistorted speech) from noisy speech (distorted speech) by converting the statistical models of noisy speech into that of clean speech without the prior knowledge on characteristics and distributions of noise source. Presently, this approach has still not attracted many researchers to apply in general noisy speech enhancement because of some major problems: those are the difficulties of noise adaptation and the lack of noise robust synthesizable features in different noisy environments. In this paper, we adopted the methods of state-of-the-art voice conversions and speaker adaptation in speech recognition to the proposed speech recovery approach applied in different kinds of noisy environment, especially in adverse environments with joint compensation of additive and convolutive noises. We proposed to use the decorrelated wavelet packet coefficients as a low-dimensional robust synthesizable feature under noisy environments. We also proposed a noise adaptation for speech recovery with the eigennoise similar to the eigenvoice in voice conversion. The experimental results showed that the proposed approach highly outperformed traditional nonlearning-based approaches.

1. Introduction

Speech is the most common information in telecommunication systems. Therefore, speech processing has been considered by numeral researchers. Quality and intelligibility of speech are degraded by different distortion sources such as background noises, commonly assumed as additive noises, channel noise, commonly assumed as convolutive noises, and distortion caused by speech disorders. Thus, clean (or undistorted) speech recovery is critical for speech communications.

Present single microphone noisy speech enhancement algorithms have been efficiently used for additive noise but inefficient for convolutive noise because only additive noise can be easily modeled as an independent Gaussian noise [1–4]. Moreover, quality and intelligibility of speech are greatly degraded in adverse environments with joint compensation of additive and convolutive noises, but there is still a lack of efficient methods to solve this problem.

Although multimicrophone models outperform single models [5], the requirement of having more than one microphone is not always practical. Therefore, developing a method for speech recovery in both additive and convolutive noises environments, especially in joint compensation of additive and convolutive noises, when only one microphone source is provided, is a critical and interesting research topic.

In the literature, there are a few researches on learning-based speech enhancement [6–13]. Among them, learning-based speech enhancement approach using statistical spectral conversion has been proposed for alaryngeal speech caused by speech disorders [8, 9], body-conducted speech [10], NAM-captured speech [11], and bone-conducted speech [12]. This approach is adapted from the concept of voice conversion that can be applied for both additive and convolutive noises using only single microphone. This approach can be also applied for other kinds of distortion as in speech disorders. Therefore, it might be a general learning-based approach used for speech enhancement with all kinds of

distortions. However, general learning-based approaches also have still not attracted many researchers in the field of noisy speech enhancement, due to two main problems; those are the inefficiency of adaptation techniques and the lack of a low-dimensional robust synthesizable speech features for different noisy environments.

In this paper, we proposed a learning-based noisy speech enhancement approach that we call "eigennoise" approach, adopted from the terms "eigenface" in face recognition [14] and "eigenvoice" in voice conversion [15]. In the proposed approach, we solved the two drawbacks of learning-based speech enhancement approach using spectral conversion, in which we proposed a low-dimensional robust synthesizable wavelet-based feature, and a noise-independent modeling combined with a noise adaptation method. We evaluated the proposed method with other spectral-conversion-based methods and other traditional nonlearning-based methods with different kinds of noise, including additive noise, convolutive noise, and joint compensation of additive and convolutive noise, with SNR from ultra-low to high. The experimental results showed that the proposed approach highly outperformed traditional nonlearning-based approaches.

This paper is organized as follows. In Section 2, brief on noise modeling in speech is described; Section 3 presents the GMM-based statistical spectral conversion that we use for the proposed noisy speech enhancement approach; in Section 4, wavelet-based robust and synthesizable speech features that we used for the proposed noisy speech enhancement method are described. The generalized learning-based speech enhancement using spectral conversion approach is described and discussed in Section 5. Finally, our work is summarized in the last section.

2. Noise Modeling in Speech

Noisy environment can be modeled by a background noise $b(n)$ or/and a distortion channel $h(n)$. In the ideal case, background noises are supposed additive while distortion channels are convolutive [16].

Assume that the clean speech is $s(n)$ and the noisy speech is $x(n)$. In the ideal case with a convolutive channel noise source $h(n)$, the noisy speech is determined as in (1) and Figure 1(a):

$$x(n) = s(n) * h(n). \tag{1}$$

In the ideal case with a additive background noise source $b(n)$, the noisy speech is determined as in (2) and Figure 1(b):

$$x(n) = s(n) + b(n). \tag{2}$$

Real noises can be compensated from both background noise $b(n)$ and channel noise $h(n)$, and the noisy speech can be modeled as in (3), (4), and Figures 1(c) and 1(d):

$$x(n) = s(n) * h(n) + b(n) \tag{3}$$

$$x(n) = (s(n) + b(n)) * h(n). \tag{4}$$

Noise also is classified into stationary and nonstationary noises. In stationary noises, the noise spectrum levels do not change over time or position. On the contrary, in nonstationary noise, the spectrum levels change in time, and it does not include a trend-like behavior. Most of researches on noise reduction and also this paper are based on the assumption that noise is stationary.

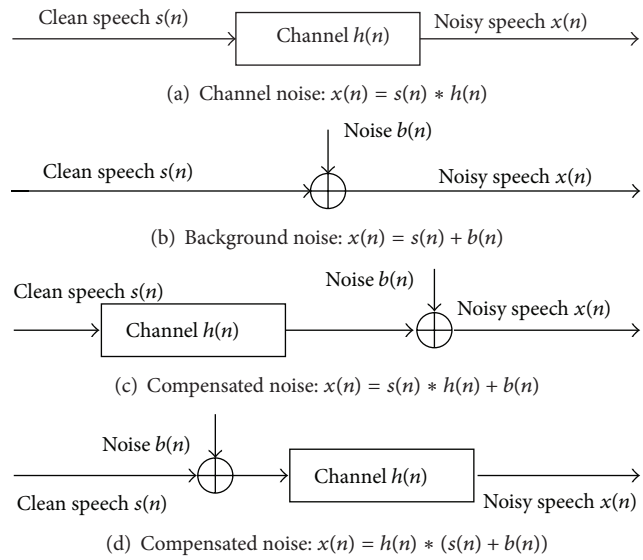

Figure 1: Artificial noise environments.

3. GMM-Based Statistical Spectral Conversion

3.1. Learning Stationary Information in Speech. In most of learning-based speech applications, the "stationary" assumption helps us to avoid the learning with big data. An example is given for speaker recognition when speaker identity is characterized by the variation of short-time spectral parameters resulting in that speaker identity can be recognized by an unsupervised short training and short testing utterances [17]. This approach is based on the fact that the speaker individual is "stationary" information that is fully represented in short utterances. With an assumption on "stationary" characteristics, any applications can be trained and recognized by a short training and short testing utterances.

In this section, we review some popular learning methods that can be used for learning different kinds of "stationary" information of speech with a short training: those are the neural network, the HMM, and the GMM.

Rosenblatt [18] developed the "perceptron," which was modeled after neurons. It is the starting point for numeral later works on neural networks. The performance of neural networks has been improved far from the starting point. However, neural networks have still high computational cost compared with statistical learning methods.

Statistical machine learning has been proposed with many advantages compared with neural networks [19]. There are many statistical machine learning methods and algorithms. The two most popular statistical methods used for speech applications are the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM). The probabilistic

HMM modeling is suitable for text-dependent speech applications such as speech recognition/synthesis [20]. However, in text-independent speech applications such as speaker recognition or spectral conversion, the sequencing of sounds found in the training data does not necessarily reflect the sound sequences found in the testing data [21]. This is also supported by experimental results in [22] which found text-independent performance was unaffected by discarding transition probabilities in HMM models.

Therefore, GMM might be one of the most suitable learning methods for training with big speech data in text-independent applications such as noisy speech enhancement. In this paper, GMM is used for training, integrated with a sparse low-dimensional speech feature.

3.2. GMM-Learning in Spectral Conversion. As mentioned in previous section, GMM seems one of the most efficient statistical learning methods for training with speech data in text-independent speech applications. GMM is also the most popular training method used in spectral conversions [15, 21]. In this subsection, we present briefly the training and predicting procedure using GMM-based statistical voice conversion that we used for the proposed noisy speech enhancement method.

3.2.1. Training Procedure. The time-aligned source feature is represented by a time sequence $X = [X_1^T, X_2^T, \ldots, X_N^T]$, where N is the number of frames. The time-aligned target feature is represented by a time sequence $Y = [Y_1^T, Y_2^T, \ldots, Y_N^T]$, where X_n and Y_n are the D-dimensional feature vectors for the nth frame. Using parallel training dataset consisting of time-aligned source and target features $[X_1^T, Y_1^T], \ldots, [X_N^T, Y_N^T]$, where T denotes transposition of the vector, a GMM on joint probability density $p(X, Y \mid \lambda)$ is trained in advance as follows:

$$\lambda = \arg\max \prod_{n=1}^{N} p\left(X_n, Y_n \mid \lambda\right), \tag{5}$$

where λ denotes model parameters. The joint probability density is written as

$$\begin{aligned} p\left(X_n, Y_n \mid \lambda\right) &= \sum_{i=1}^{M} \alpha_i N\left(X_n, Y_n; \mu_i^{(X,Y)}, \Sigma_i^{(X,Y)}\right), \\ \mu_i^{(X,Y)} &= \begin{bmatrix} \mu_i^{(X)} \\ \mu_i^{(Y)} \end{bmatrix}, \\ \Sigma_i^{(X,Y)} &= \begin{bmatrix} \Sigma_i^{(XX)} & \Sigma_i^{(XY)} \\ \Sigma_i^{(YX)} & \Sigma_i^{(YY)} \end{bmatrix}, \end{aligned} \tag{6}$$

where M is the number of Gaussian mixtures. $N(x; \mu_i, \Sigma_i)$ denotes the 2D dimensional normal distribution x with the mean μ_i and the covariance matrix Σ_i.

The ith mixture weight is α_i which is the prior probability of the joint vector $[X^T, Y^T]$ and it satisfies $0 \leq \alpha_i \leq 1$, $\sum_{i=1}^{M} \alpha_t = 1$. The parameters $(\alpha_i, \mu_i, \Sigma_i)$ for the joint density $p(X, Y \mid \lambda)$ can be estimated using the expectation maximization (EM) algorithm.

3.2.2. Conversion Procedure. The transformation function that converts source feature X to target feature Y is based on the maximization of the following likelihood function:

$$p\left(Y \mid X, \lambda\right) = \sum_{m} p\left(m \mid X, \lambda\right) p\left(Y \mid X, m, \lambda\right), \tag{7}$$

where $m = \{m_{i1}, m_{i2}, \ldots, m_{iN}\}$ is a mixture sequence.

At frame nth, $p(m_i \mid X_n, \lambda)$ and $p(Y_n \mid X_n, m_i, \lambda)$ are given by

$$p\left(m_i \mid X_n, \lambda\right) = \frac{\omega_i N\left(X_n; \mu_i^{(X)}, \Sigma_i^{(XX)}\right)}{\sum_{j=1}^{M} \omega_j N\left(X_n; \mu_j^{(X)}, \Sigma_j^{(XX)}\right)}, \tag{8}$$

$$p\left(Y_n \mid X_n, m_i, \lambda\right) = N\left(Y_n; E_n\left(m_i\right), D\left(m_i\right)\right), \tag{9}$$

where

$$\begin{aligned} E_n\left(m_i\right) &= \mu_i^{(Y)} + \Sigma_i^{(YX)} \Sigma_i^{(XX)-1}\left(X_n - \mu_i^{(X)}\right), \\ D\left(m_i\right) &= \Sigma_i^{(YY)} - \Sigma_i^{(YX)} \Sigma_i^{(XX)-1} \Sigma_i^{(XY)}. \end{aligned} \tag{10}$$

A time sequence of the converted feature $\widehat{y} = [\widehat{y}_1^T, \widehat{y}_2^T, \ldots, \widehat{y}_N^T]^T$ is computed as follows:

$$\widehat{y} = \arg\max p\left(Y \mid X, \lambda\right). \tag{11}$$

The converted features can be estimated by the EM algorithm.

3.2.3. Universal Background Model. There are one-to-one, many-to-one, and one-to-many VC systems [15]. In many-to-one VC, the full training with all sources is expensive and sometime impossible, similar to the full training with all targets in one-to-many VC. Therefore, the source (target) independent model called UBM is introduced in the GMM-UBM speaker verification system [16], while a single, speaker-independent background model is used. The UBM is a large GMM trained to represent the speaker-independent distribution of features. The independent UBM is then used as a representative target with one-to-many VC and a representative source with many-to-one VC.

There are two main approaches that can be used to obtain the UBM model. The first approach is to simply pool all the data to train the UBM via the EM algorithm (Figure 2(a)). The second approach is to train individual UBMs over the subpopulations in the data and then pool the subpopulation models together (Figure 2(b)). In this paper, the first approach is used to train GMM-UBM due to its simplicity. When using this approach, training procedure is the same as presented in Section 3.2.1 but the training data is combined from many noisy speech conditions and environments.

3.2.4. MAP Adaption. Using UBM-GMM is useful with one-to-many and many-to-one VCs. However, to improve

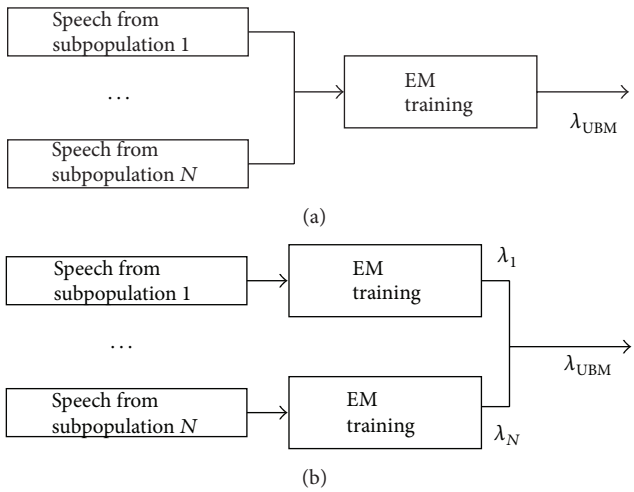

Figure 2: GMM-UBM training: (a) data from subpopulations are pooled prior to training the UBM via EM algorithm; (b) individual subpopulation models are trained then pooled to final UBM model.

the estimation of the model, the maximum a posteriori (MAP) adaption, also known as Bayesian learning estimation, is used to adapt the UBM into the required source (many-to-one) or target (one-to-many) models [16]. In the proposed noisy speech enhancement framework, MAP adaption is used to adapt noise-independent model to the models of specific noisy conditions.

Although all weights, means, and variances of GMM can be adapted using MAP, experiments show that only adapting the mean of GMM obtained best performance [16]. The MAP adaption for the mean of GMM is represented as below.

Given a UBM and training vectors, $X = [x_1, x_2, \ldots, x_T]$, we first determine the probabilistic alignment of the training vectors into the UBM mixture components. That is, for mixture ith in the UBM, we compute

$$\Pr(i \mid x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)}. \tag{12}$$

We then use $\Pr(i \mid x_t)$ and x_t to compute the sufficient statistics for the mean parameter:

$$\begin{aligned} n_i &= \sum_{t=1}^{T} \Pr(i \mid x_t), \\ E_i(x) &= \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid x_t)\, x_t. \end{aligned} \tag{13}$$

Finally, these new sufficient statistics from the training data are used to update the old UBM sufficient statistics for mixture i to create the adapted parameters for mixture ith with

$$\widehat{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m)\, \mu_i. \tag{14}$$

The adaptation coefficients controlling the balance between old and new estimates are α_i^m.

4. Noise Robust Synthesizable Wavelet-Based Features

4.1. Noise Robustness of Traditional Speech Features. It is known that noisy speech signal varies largely with different kinds of noise. In learning-based speech applications, it is able to recognize and synthesize perfectly noisy speech as in clean environment if the noisy environment of training data is identical to that of the testing data. Unfortunately, the noisy environment of testing data is seldom known in advance, and it is difficult to train the data in all possible kinds of noisy environment. When the noisy environments of training data are different from the noisy environment of testing data, the recognition and synthesis system typically performs much worse. Therefore, it is necessary to understand and eliminate variance in the speech signal due to the environmental changes and thus ultimately avoid the need for extensive training in different noisy environments. As a consequence, the learning-based speech applications in noisy environments require robust features, which are insensitive with noise environments.

State-of-the-art speech recognition is based on source/filter model to extract vocal tract features or spectral envelope features separated with source features. The two most popular spectral envelope features are linear prediction coefficients (LPC), and Mel-Frequency-Cepstral-Coefficients (MFCC). The most popular source feature is fundamental frequency ($F0$).

While LPC has been shown to be sensitive with noises, MFCC is robust with noisy environments and it is the state-of-the-art and standard feature for speech recognitions in both clean and noisy environments [23]. The robustness of MFCC is mostly caused by the perceptual Mel-scale, integrated inside MFCC. The nonlinear Mel-scale follows the psychoacoustic model which is natural with human hearing [24]. Because humans are capable of detecting the desired speech in a noisy environment without prior knowledge of the noise, modeling speech features closing with human hearing has been improved performance of speech applications in noisy environments.

However, MFCC is built based on the concept of Short Time Fourier Transform (STFT), in which fixed length window is used for analysis. The basis vectors of MFCC cover all frequency bands, so corruption of a frequency band of speech by noise affects all MFCC coefficients. Therefore, researchers still attempt to improve the noise robustness of MFCC and to propose other noise robust features for speech applications in noisy environments.

4.2. Synthesizability of Traditional Speech Features. Feature extraction is a critical analysis stage for both recognition and synthesis systems. In recognition tasks, the features can be any parameter that characterizes speech. However, in synthesis tasks, the speech features are usually required invertible or synthesizable.

LPC and MFCC features are two indirectly synthesizable features. To synthesize speech, LPC or MFCC needs to be combined with $F0$ in a VOCODER, which is one popular

source/filter synthesizer widely used in speech coding and synthesis [20].

In VOCODER, $F0$ and random noise are used for source excitation. In the literature, many researches show that VOCODER produces "buzzy" synthetic speech [25]. Therefore, the requirement of combination with $F0$, which causes the indirect synthesizability of MFCC, limits the efficiency of MFCC in speech synthesis. Currently, MLSA filter, one kind of VOCODER using MFCC and $F0$, has been still used in state-of-the-art TTS [20]. The use of directly synthesizable features without using source/filter model is expected to solve the "buzzy" problem of VOCODER in speech synthesis.

4.3. Perceptual Wavelet Packet. The discrete wavelet transform (DWT) of a general continuous signal $s(t)$ is the family $C(a,b)$, defined in (15):

$$\begin{aligned} C(a,b) &= \int_R s(t)\,\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)dt, \\ a &= 2^j, \\ b &= k2^j, \\ &\qquad (j,k) \in Z^2. \end{aligned} \tag{15}$$

The indexes of $C(a,b)$ are called wavelet coefficients.

The inverse DWT (IDWT) reconstructs the signal $s(t)$ from wavelet coefficients $C(a,b)$ as in (16):

$$s(t) = \sum_{j\in Z}\sum_{k\in Z} C(j,k)\,\psi_{j,k}(t). \tag{16}$$

With the deep mathematical formulation of DWT, IDWT can be found in [26].

In wavelet analysis (wavelet decomposition), using the DWT, a signal is split into an approximation and a detail. The approximation is then itself split into a second-level approximation and detail, and the process is repeated. Wavelet synthesis (wavelet reconstruction) is the inverse process of the wavelet analysis.

In wavelet packet analysis, the details as well as the approximations can be split; therefore, the subband structure can be customized with a user-defined wavelet tree. Recent researches in wavelet show that the integration of the wavelet packet and the psychoacoustic model into the perceptual wavelet packet transform (PWPT) may improve performance of speech applications [27–29].

It is shown in the literature that PWPT has significant performances with noisy speech recognition and speech coding in comparison to the conventional wavelet.

In psychoacoustic model, frequency components of sounds can be integrated into critical bands that refer to bandwidths at which subjective response become significantly different. One widely used critical band scale is Mel-scale in MFCC [24], and another popular scale is Bark scale [30].

The Mel-scale m can be approximately expressed in terms of the linear frequency as in

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) = 1127\log_e\left(1+\frac{f}{700}\right). \tag{17}$$

The Bark scale z is approximately expressed as

$$\begin{aligned} z(f) &= 13 * \arctan\left(7.6 * 10^{-4} f\right) + 3.5 \\ &\quad * \arctan\left(1.33 * 10^{-4} f\right)^2. \end{aligned} \tag{18}$$

In (17), (18) f is the linear frequency in Hertz.

The nonlinear Mel-scale and Bark scale can be used to design the wavelet tree for the perceptual wavelet packet. PWPT has been used to extract robust and synthesizable features in the literature.

4.4. Noise Robustness and Synthesizability of Wavelet-Based Features. Wavelet has fine time and frequency resolution, and the effects of noise on speech are localized in some specific subbands. Therefore, wavelet is expected to be an efficient tool for noise robust feature extraction.

In the literature, there are many noise robust wavelet speech features that have been proposed. These features can be grouped into two main categories.

The first category computes the sum (or weighted sum) of energies in each subband to form the whole feature [27, 28].

The second category simply uses the wavelet coefficients, retaining the time information, to form the feature [29]. There are also some mixed categories.

In the first category, the time information in the wavelet subbands is lost into the subband energies. Moreover, this kind of features is noninvertible or nonsynthesizable.

The use of wavelet coefficients in the second category is simple while keeping noise robustness of the wavelet analysis. Moreover, it is known that inverse wavelet transform can perfectly reconstruct signal from wavelet coefficients. Therefore, the feature using wavelet coefficients is robust with noise and synthesizable and this feature is used in the proposed noisy speech enhancement method in this paper.

4.5. Features Decorrelation and Compression with DCT. Although the feature using wavelet coefficients is robust with noise and synthesizable, the simple feature concatenated from all wavelet coefficients in all sub-bands is very high-dimensional. Moreover, wavelet coefficients are correlated within and between sub-bands [31]. Therefore, if we want to use wavelet coefficients feature in both recognition and synthesis tasks, we need to decorrelate and compress wavelet coefficients feature vector. Although principal component analysis (PCA) [32] is the most popular method to reduce feature dimension, it is difficult to reconstruct original feature for synthesis tasks. The most popular invertible decorrelation transform is DCT [33].

The most common DCT definition of a 1D sequence of length N [33] is

$$C(u) = \alpha(u)\sum_{x=0}^{N-1} f(x)\cos\left[\frac{\pi(2x+1)u}{2N}\right] \tag{19}$$

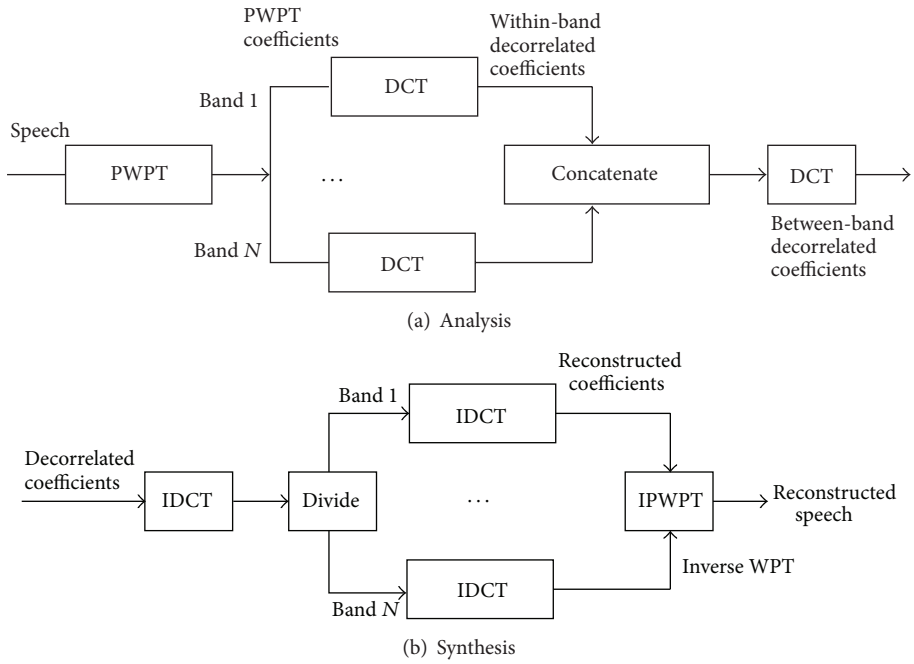

Figure 3: Feature extraction and reconstructions.

for $u = 0, 1, 2, \ldots, N-1$. Similarly, the inverse transformation is defined as

$$f(x) = \sum_{u=0}^{N-1} \alpha(u) C(u) \cos\left[\frac{\pi(2x+1)u}{2N}\right] \quad (20)$$

for $x = 0, 1, 2, \ldots, N-1$. $\alpha(u)$ is defined as

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0, \\ \sqrt{\frac{2}{N}}, & u \neq 0. \end{cases} \quad (21)$$

The DCT is often used in signal processing because it has a strong "energy compaction" property [34] where most of the signal information tends to be concentrated in a few low-frequency components of the DCT. In this paper, DCT is used to decorrelate wavelet-based feature in this proposed noisy enhancement method.

5. Eigennoise Speech Recovery Framework

The approach of using eigenfaces for recognition was developed by Sirovich and Kirby [14]. A set of eigenfaces can be generated by PCA on a large set of images depicting different human faces. Informally, eigenfaces can be considered a set of "standardized face ingredients," derived from statistical analysis of many pictures of faces.

Developed from the concept of eigenfaces, Ohtani et al. proposed an eigenvoice-GMM voice conversion [15].

In this paper, we call the UBM of a large set of noisy environments "eigennoise," and then we proposed a speech recovery approach using GMM-UBM-MAP based on the joint factor analysis that we called "eigennoise" approach.

The noise robust synthesizable feature analysis and synthesis are described in Figure 3, and we use PWPT to extract wavelet coefficients from input speech. The coefficients in each subband are highly correlated and bear a lot of redundant information, especially that, in high bands, there are a lot of small coefficients or zeros. Therefore, DCT is used in each subband to decorrelate within-band correlations. After concatenating coefficients from all bands to form the whole coefficients, DCT is used again to decorrelate the between-band correlations. Both PWPT and DCT are completely invertible; thus, the speech is perfectly reconstructed.

The eigennoise training and conversion are presented in Figure 4. Noisy speech with several noisy environments is used for training the noise-independent eigennoise model (GMM-UBM) as shown in Figure 4(a) and presented in Section 3.2.3. The noise-independent model is then adapted to each noise-dependent noise model later as shown in Figure 4(b) and presented in Section 3.2.4. Clean speech is converted from correspondent noisy speech and noise-dependent model as shown in Figure 4(c) and presented in Section 3.2.2.

6. Implementation and Evaluations

6.1. Data Preparation. The clean speech data used in our evaluation was the well-known English MOCHA-TIMIT. The noise database is the NOISEX-92. We created some artificial noisy environments simulating the additive background noise, the convolutive channel noise, and the mixed noise. The noise sources were selected from NOISEX-92.

We simulated the practical open-dataset testing, in which the testing noisy condition does not match the training conditions. The noise inputs of the artificial noisy environment

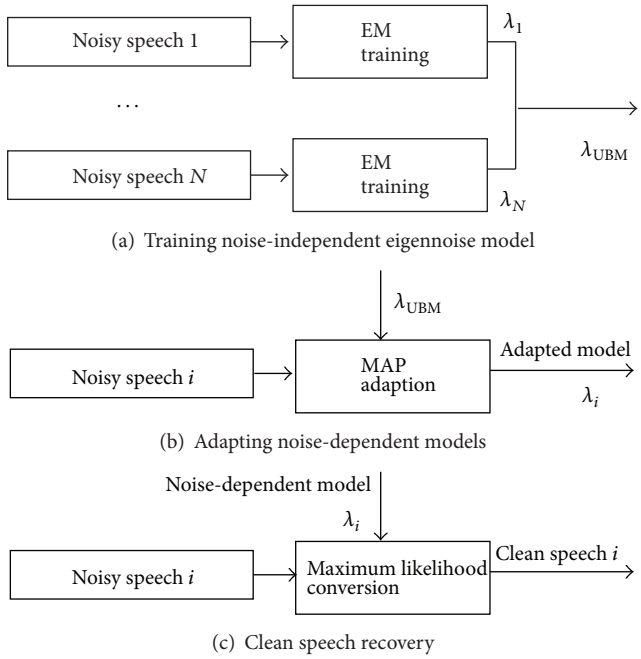

Figure 4: Eigennoise training and conversion.

used for training and testing were factory noise. The signal-to-noise ratios (SNRs) of the noisy speech used for training were −5, 5, 15, 25, and 35 dB, but those used for testing were −10, 0, 10, 20, and 30 dB.

All enhanced noisy speech was evaluated with the objective tests while only mixed noisy speech with SNR of the noisy speech approximated −10 dB, which was the closest with the ultra-high real noises, was evaluated with the subjective test.

6.2. Implementation Parameters. In all experiments, the number of Gaussian components M, which should be chosen large enough if we have enough data for training, was chosen as 15. The adaptation coefficient α was initially set to 0.5. The speech data used for all evaluations were resampled at 8 KHz, yielding a bandwidth of 4 KHz, and there are approximately 17 first critical bands among 25 bands in Bark scale. Therefore, the wavelet coefficients feature had 17 coefficients. The order of LP analysis P was chosen as 17 also. The frame size was chosen as 30 ms and the overlapped interval was 15 ms.

6.3. Objective Evaluations for Speech Quality. To evaluate our proposed "eigennoise" approach with two features LP [13] and wavelet, we implemented and compared our methods with the standard nonlearning-based spectral subtraction [4] and Wiener-filter [5] methods. We used Peak-Signal-Noise ratio (PSNR) for objective evaluation. The average PSNR results are shown in Figures 5(a), 5(b), and 5(c). The results reveal that while quality of enhanced speech with nonlearning-based methods depended linearly on the input SNRs, it was acceptable with additive noises but very bad with convolutive and mixed noises. On the contrary, performance of learning-based methods was quite independent with the input SNRs, as well as the kinds of noise. The proposed "eigennoise" speech recovery with wavelet-GMM outperformed the proposed "eigennoise" speech recovery with LP-GMM. In general, learning-based noisy speech enhancements greatly outperformed the nonlearning-based methods.

6.4. Subjective Evaluations for Speech Intelligibility. The speech signals of 100 English words with clean, noisy, and enhanced signals were played in random order in the tests for 5 native English subjects. The subjects were asked to listen to each word only once and write down what they heard. Speech intelligibility could generally be evaluated using the average recognition accuracy scored by all subjects. The results are shown in Figure 6. The subjective evaluation results also support that the nonlearning-based methods reduced the intelligibility of speech while the learning-based methods improved much intelligibility of speech. In addition, the "eigennoise" method with wavelet-GMM outperformed that with LP-GMM method.

7. Conclusions and Discussions

For adverse environments with joint compensation of additive and convolutive noises, one of the biggest challenges in noisy speech enhancement, the proposed learning-based approach using spectral conversion presented in this paper, is one promising candidate among a few available approaches.

However, the proposed framework and methods have still some remaining issues needed to be studied in the future. The two biggest issues are the requirements of hardware performance for training and the efficient training methods with big data.

There are many kinds of real noise. Thus, to build a practical learning-based noisy speech enhancement usually requires training with several noisy speech conditions, corresponding with the several real noise environments. Therefore, the hardware performance is required to be very high to cope with training of huge corpus. This requirement is not available at the turn of the millennium. There is good news that the hardware performance has been developed rapidly recently. While the first processor Intel 8080 has a clock rate of 2 MHz, speed of modern present processor overcomes 8 GHz [35]. In addition, processing performance of computers is increased by using multicore processors, in which processor can handle numerous asynchronous events, interrupts, and so forth. Recent advances in computer hardware researches and applications reduce the difficulty of training with gigantic corpus. Therefore, the first limitation of the learning-based noisy speech enhancement can be overcome with newest hardware technologies.

With the rapid developments of statistical learning methods, learning-based noisy speech enhancement could be much more efficient compared with the beginning results. In this paper, we proposed a wavelet-GMM method that we call eigennoise speech recovery method. The GMM-based training methods have been shown to be efficient with big speech data. However, the computational cost is still necessary to be improved in the future.

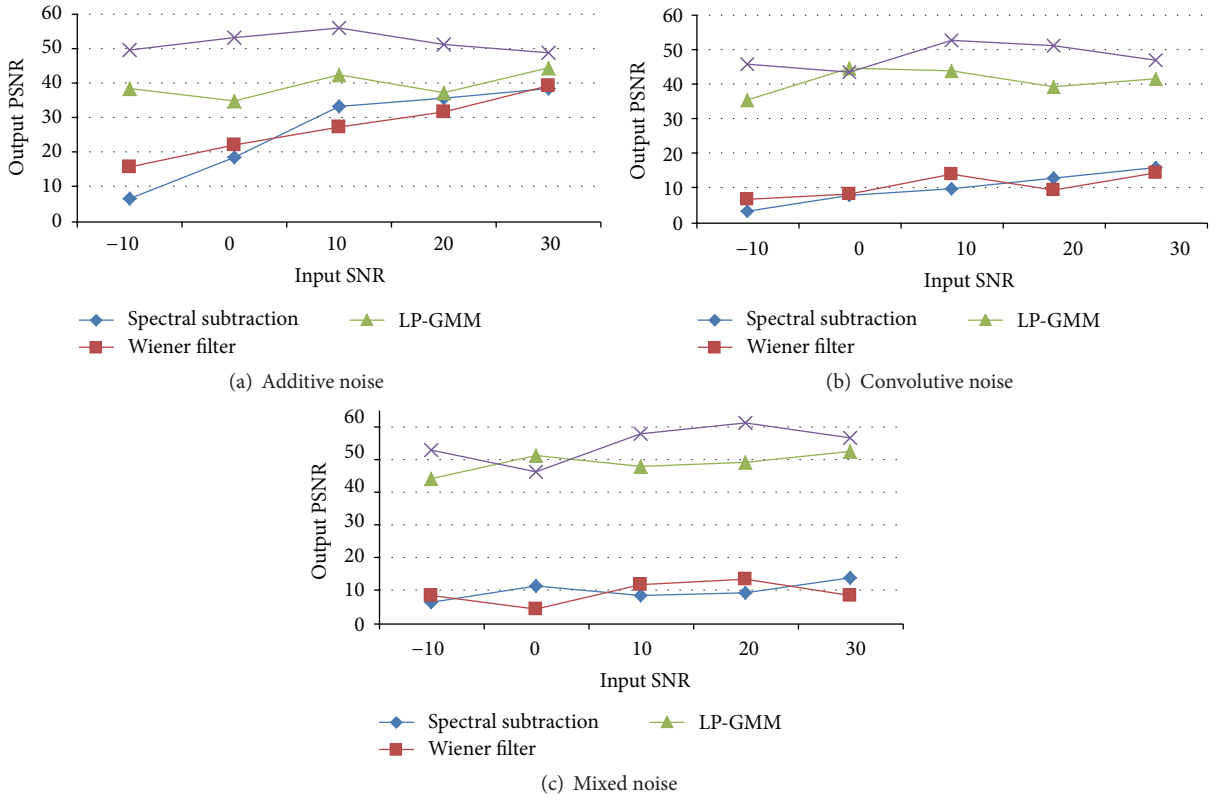

Figure 5: Objective evaluations: (a) additive noise, (b) convolutive noise, and (c) mixed noise.

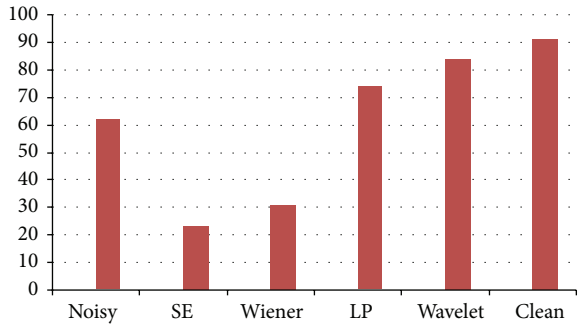

Figure 6: Subjective evaluations.

One other disadvantage of the proposed methods is the speaker-dependent requirement not mentioned in this paper. In the future, we will also compare deep neural models with the proposed model and evaluate the proposed method with a noisy speech recognition system to confirm the efficiency of the proposed model.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[6] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in Neural Information Processing Systems*, vol. 13, pp. 758–764, 2001.

[7] A. Mouchtaris, J. V. Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single-channel speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1180–1193, 2007.

[8] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 2, pp. 97–105, 1997.

[9] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[10] T. Hirahara, M. Otani, S. Shimizu et al., "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Communication*, vol. 52, no. 4, pp. 301–313, 2010.

[11] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[12] T. N. Phung, M. Unoki, and M. Akagi, "Improving bone-conducted speech restoration in noisy environment based on LP scheme," in *Proceedings of the APSIPA Annual Summit and Conference*, Singapore, December 2010.

[13] D. Huy-Khoi, P. Trung-Nghia, H. C. Nguyen, V. T. Nguyen, and Q. V. Thai, "A novel spectral conversion based approach for noisy speech enhancement," *International Journal of Information and Electronics Engineering*, vol. 1, no. 3, pp. 281–285, 2011.

[14] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of the Optical Society of America A. Optics and Image Science*, vol. 4, no. 3, pp. 519–524, 1987.

[15] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Transactions on Information and Systems*, vol. 93, no. 6, pp. 1589–1598, 2010.

[16] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 975–983, 2005.

[17] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[18] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[19] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[20] H. Zen, T. Nose, J. Yamagishi et al., "The HMM-based speech synthesis system version 2.0," in *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW '07)*, Bonn, Germany, August 2007.

[21] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 285–288, IEEE, May 1998.

[22] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 2, pp. 157–160, IEEE, San Francisco, Calif, USA, March 1992.

[23] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Technical standard 201 108, v1.1.3, 2003.

[24] S. S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[25] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 2, pp. 861–864, IEEE, Seattle, Wash, USA, May 1998.

[26] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley Cambridge Press, Wellesley, Mass, USA, 1997.

[27] Q. T. Nguyen and T. N. Phung, "The perceptual wavelet feature for noise robust Vietnamese speech recognition," in *Proceedings of the 2nd International Conference on Communications and Electronics (HUT-ICCE '08)*, pp. 258–261, Hanoi, Vietnam, June 2008.

[28] P. T. Trung-Nghia, D. D. Cuong, and P. V. Binh, "A new wavelet-based wide-band speech coder," in *Proceedings of the International Conference on Advanced Technologies for Communications (ATC '08)*, pp. 349–352, Hanoi, Vietnam, October 2008.

[29] R. C. Guido, L. Sasso Vieira, S. Barbon Júnior et al., "A neural-wavelet architecture for voice conversion," *Neurocomputing*, vol. 71, no. 1–3, pp. 174–180, 2007.

[30] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, article 248, 1961.

[31] P. F. Craigmile and D. B. Percival, "Asymptotic decorrelation of between-scale wavelet coefficients," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 1039–1048, 2005.

[32] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[33] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 90–93, 1974.

[34] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, Boston, Mass, USA, 1990.

[35] M. Chiappetta, "AMD breaks 8GHz overclock with upcoming FX processor, sets world record," *Hot Hardware*, 2011.