J Sign Process Syst (2016) 82:141–150
DOI 10.1007/s11265-015-1025-1

Speech Enhancement Based on Analysis–Synthesis Framework with Improved Parameter Domain Enhancement

Bin Liu[1] · **Jianhua Tao**[1] · **Zhengqi Wen**[1] · **Fuyuan Mo**[2]

Received: 4 May 2015 / Revised: 18 June 2015 / Accepted: 6 July 2015 / Published online: 24 July 2015

Abstract This paper presents a speech enhancement approach based on analysis–synthesis framework. An improved multi-band summary correlogram (MBSC) algorithm is proposed for pitch estimation and voiced/unvoiced (V/UV) detection. The proposed pitch detection algorithm achieves a lower pitch detection error compared with the reference algorithm. The denoising autoencoder (DAE) is applied to enhance the line spectrum frequencies (LSFs). The reconstruction loss could be decreased compare with the swallow model. The proposed approach is evaluated using the perceptual evaluation of speech quality (PESQ) and the experimental results show that the proposed approach improves the performance of speech enhancement compared with the conventional speech enhancement approach. In addition, it could be applied to parametric speech coding even at low bit rate and low signal-noise ratio (SNR) environments.

Keywords Analysis-synthesis framework · Multi-band summary correlogram · Denoising autoencoder · Speech enhancement · Speech coding

✉ Bin Liu
liubin@nlpr.ia.ac.cn

Jianhua Tao
jhtao@nlpr.ia.ac.cn

Zhengqi Wen
zqwen@nlpr.ia.ac.cn

Fuyuan Mo
mofuyuan@aliyun.com

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2] Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

1 Introduction

Single-channel speech enhancement is an important branch of speech signal processing. It is useful for many applications such as speech recognition, speech coding and hearing aid; it is often used as a pre-processer to improve speech quality and intelligence. Enhancing the speech signal from noise corrupted signal had been addressed as a challenging topic. It is particularly difficult to track the non-stationary noise especially in low signal-noise ratio (SNR) environments from a single -channel noisy speech.

A large number of speech enhancement approaches have been proposed such as spectral subtraction [1] and minimum mean square error (MMSE) [2]. A major drawback of the spectral subtraction approach is the introduction of musical noise in the enhanced speech [3]. The MMSE uses a Bayesian approach to estimate the clean speech magnitude spectrum assuming Gaussian distributions [2] or super-Gaussian distributions [4] for the speech and noise magnitudes. It has been suggested that the good performance can be attributed to the use of the decision-directed approach for estimation of the a priori SNR [5]. The mentioned above approaches focused on the analyzing the statistical difference between speech and noise. The background noise is reduced frame by frame. The correlation between adjacent frames is not considered effectively. The supervised enhancement approaches have been shown to produce better quality speech signals compared with the unsupervised approaches. Some examples of supervised algorithms include the codebook-based [6], hidden Markov model (HMM) [7] and artificial neural network (ANN) [8]. The neural network is suitable for modeling the non-linear relationship between clean speech and noisy speech. The shallow neural network (SNN) as nonlinear mapping has also been proposed [8]. Nevertheless, the SNN starts from random initialization and often leads to local minima or plateaus [9].

In [10], an analysis–synthesis framework is proposed to resynthesize clean speech signals based on acoustic parameters (pitch, spectral gain and spectral envelope) extracted from noisy speech. The target speech is reconstructed with related acoustic parameters only and background noise is automatically removed. This approach is attractive due to it can retrieve damaged harmonic structure and at the same time eliminates musical noise. However, to ensure accurate model parameter estimation, a preprocessing step is often required to preclean the noisy signals prior to speech enhancement based on analysis–synthesis framework. It is reported in [10] that pitch and spectral gain estimation applied on precleaned spectrum can give satisfactory result even in very low SNR environments. However, most precleaning algorithms are difficult to recover the spectral envelope which has been distorted by background noise. Both pitch estimation and spectrum envelope estimation are important to improve the performance for speech enhancement.

Pitch estimation algorithms can be broadly classified into three categories: time-domain, frequency-domain, and time frequency- domain. Time-domain F0 estimation directly exploit a signal's temporal periodicity [11]. Frequency-domain F0 estimation make use of the signal's short-time spectral harmonicity [12]. Time-frequency domain F0 estimation often separates a signal into various frequency bands, and then applies time-domain processing in each frequency band [13]. The auditory-model correlogram-based F0 estimation is a popular time-frequency domain method. It can yield estimates close to human's perceived pitch for signals and also have the potential to be noise-robust. An SNR-weighted correlogram based F0 estimation using multi-band comb filters is proposed in [14]. The proposed F0 estimation is effective in improving the accuracy of F0 estimation in the presence of noise. The subband which has high voicing strength represents obvious harmonic structure and it is effective to pitch estimation.

The method to improve the spectral envelope estimation can be regarded as a problem of estimating spectral envelope parameters of clean speech from noisy speech. It is well known that Wiener filtering is correlated with linear prediction, and clean spectral envelope parameters can be iteratively estimated from noisy speech using Wiener filtering [6]. Besides, the Kalman filter is also widely studied in speech enhancement. In [15], it incorporates Kalman filter to track the temporal trajectories of line spectrum frequencies (LSFs). The enhanced LSFs are then directed into the analysis–synthesis framework to improve the spectral envelope estimation, and hence the performance of speech enhancement. The Gaussian mixture model (GMM) is a typical mapping model which is widely applied to voice conversion [16] and artificial wideband extension [17]. It is effective to reconstruct the target spectrum envelope. Deep learning has emerged as a new area of machine learning research [18]. It can discover the underlying regularity of multiple features, and have strong generalization abilities than shallow models. The basic strategy is to train a deep network with greedy layer wised pre-training plus fine tuning. For spectral envelope enhancement, we are paying attention to learning a mapping between the noisy spectral envelope and the clean spectral envelope. The denoising autoencoder (DAE) is an equivalent model [19]. It is trained to reconstruct the original spectral envelope from a corrupted spectral envelope. It is a multi-layer neural network structure. The model is trained to minimize the reconstruction loss between the original spectral envelope and the reconstructed spectral envelope.

In this paper, we present a speech enhancement approach based on analysis–synthesis framework to enhance noisy speech signals. A preliminary precleaning step is required to preclean the noisy signals. The goal of the precleaning step is to filter the noisy signals such that it is more suitable for the analysis–synthesis framework. An improved multi-band summary correlogram (MBSC) pitch detection algorithm is proposed. This work is an extension of the algorithm proposed in [20]. To improve the noise-robustness of V/UV detection and pitch estimation, the subband which has the high voicing strength is selected and the linear prediction residual signal is considered. The denoising autoencoder (DAE) is applied to build the mapping relationship between pre-cleaned LSFs and clean LSFs. The enhanced LSFs are then directed into the analysis–synthesis framework to improve the spectral envelope estimation. The proposed algorithm takes advantage of the analysis–synthesis framework to effectively eliminate musical noise. On the other hand, it looks for the mapping relationship to obtain enhanced spectral envelope through deep layer network. The noise could be suppressed and the distorted speech could be restored effectively.

The rest of this paper is organized as follows: the detail of proposed approach is introduced in section 2. The evaluation results are showed in the section 3. Conclusion will be elaborated in the section 4.

2 Proposed Algorithm

In this section, we firstly introduce the framework of the proposed speech enhancement approach. Subsequently, the further details are presented. The flowchart of proposed algorithm is shown in Fig. 1.

There are total six parts in the proposed method which includes precleaning, pitch estimation, V/UV detection, LSFs enhancement, spectral gain estimation and speech synthesis. In the precleaning stage, the logMMSE is applied to estimate the short-time spectral amplitude for each frame. The improved MBSC is applied to both pitch estimation and voiced/unvoiced detection in different

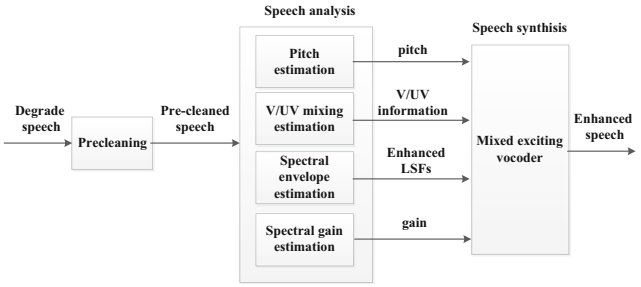

Figure 1 Block diagram of the proposed algorithm.

subband. We select the DAE model to enhance the LSFs. The spectral gain estimation is measured using a pitch adaptive window length. All estimated and enhanced speech parameters are sent to the synthesizer based on mixed excitation vocoder to reconstruct the target speech. Detailed procedures for proposed algorithm are shown as follows.

2.1 Pre-cleaning

The logMMSE is applied to precleaning. The estimation of the short-time spectral amplitude (STSA) is formulated as that of estimating the amplitude of DFT coefficients of the speech signal, given the noisy observations. The DFT coefficients of the speech signal, as well as of the noise signal are modeled as statistically independent Gaussian random variables. We are looking for the estimator, which minimized the log amplitude spectrum distortion.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively. In the time-frequency domain we have $Y(k,l)=X(k,l)+D(k,l)$, where k represents the frequency bin index and l represents the frame index. Let A denote the estimated speech amplitude

$$A(k,l) = G(k,l)\left|Y(k,l)\right| \tag{1}$$

To confirm $G(k,l)$, both speech presence and speech absence are considered. When speech is present, gain function$G_{H}1(k,l)$ *given by*

$$G_{H1}(k,l) = \frac{\sqrt{\pi}}{2}\sqrt{\frac{\xi}{\gamma(1+\xi)}}F\left[\frac{\gamma\xi}{1+\xi}\right] \tag{2}$$

with

$$F[v] = \exp\left(-\frac{v}{2}\right)\left[(1+v)I_0\left(\frac{v}{2}\right) + vI_1\left(\frac{v}{2}\right)\right] \tag{3}$$

where ξ represents priori SNR and γ is posteriori SNR; I_0 and I_1 are the modified Bessel functions of zero and first order respectively.

When speech is absent, the gain is constrained to be larger than a threshold $G_{\min}$. The gain function for MMSE is obtained by

$$G(k,l) = \{G_{H1}(k,l)\}^{p(k,l)}{G_{\min}}^{1-p(k.l)} \tag{4}$$

where $p(k,l)$ is speech presence as shown below:

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\exp\left(-\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}\right)\right\}^{-1} \tag{5}$$

where $q(k,l)$ is the a priori probability for speech absence defined by [17]

$$q(k,l) = 1-p_{local}(k,l)p_{global}(k,l)p_{frame}(k,l) \tag{6}$$

The spectral gain function is obtained with speech presence and absence. The speech presence is estimated for each frequency bin and each frame by a soft-decision, which exploits the correlation of speech presence in neighboring frequency bins and consecutive frames.

This method is superior to the MMSE STSA estimator since it results in a much lower residual noise level without further affecting the speech itself [21]. The noisy speech is initially processed by MMSE to estimate clean speech and the speech signal is reconstructed approximately. There will be large distortion in pre-cleaned speech especially for non-stationary noise. It is important to restore the distorted speech.

2.2 Pitch Estimation and Voiced/Unvoiced Detection

The block diagram in Fig. 2 gives an overview of the proposed MBSC pitch detector. This is an extension of the pitch estimation algorithm proposed in [20].

The input speech signal is first decomposed into four subbands using 32-point FIR filters. A 1-kHz filter bandwidth is chosen so that at least two harmonics are captured by each filter. When a signal contains more than 1 harmonic of the target voice speech, its envelope would typically oscillate at an amplitude modulation frequency corresponding to the inter-harmonic separation. So the Hilbert envelope in each subband is extracted. The Hilbert envelope in each subband is extracted by computing the magnitude squared of analytic signal, which is obtained by applying Hilbert transform on the FIR-filtered output. It is noted that the Hilbert envelope accurately follows the amplitude modulations of a bandpass signal. The Hilbert envelope is mean-normalized on a frame by frame basis before subsequent processing.

The lowpass-filtered non-envelope stream from the first subband is also used in subsequent processing. We consider the linear prediction residual from the first subband in subsequent processing. It is more effective for pitch detection,

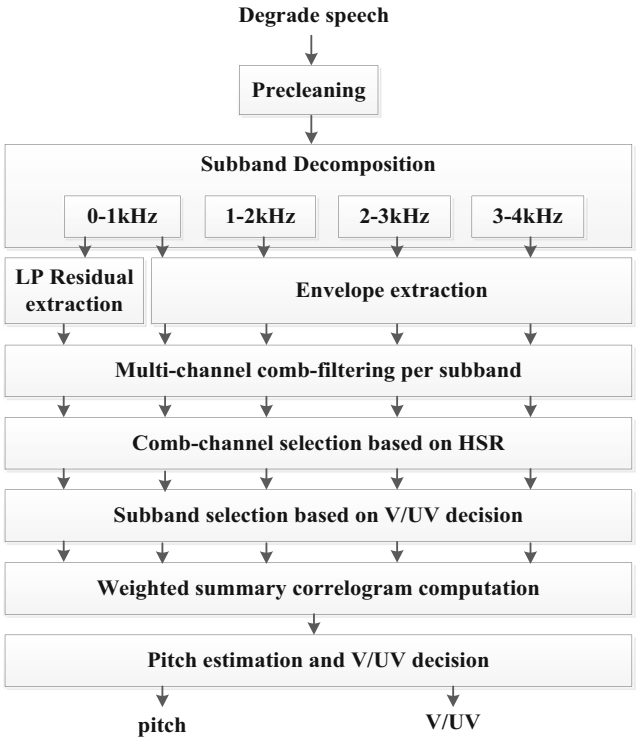

Figure 2 Block diagram of the proposed pitch detector.

especially when the first harmonic is not attenuated or noise corrupted compare with the original signal.

Multi-channel comb-filtering is performed in the frequency domain, by multiplying the input stream spectrum with a comb-function represented in the frequency domain. Multi-channel comb filtering is performed separately for each subband stream. The comb-functions are formulated using raised-cosines; this comb-function enhances spectral harmonics and suppresses the energies at the subharmonics. The raised-cosine function is selected due to its broad spectral peak lobes and smooth peak-to-valley transitions. By incorporating separate comb-filters that span different subbands, the pitch detection performance is more robust.

Harmonic-to-subharmonic energy ratio (HSR) is a measure computed to aid the selection of reliable comb-filter channels for each subband. Comb-filters defined capture the harmonic energy of the signal. To capture the subharmonic or inter-harmonic energy, inverted comb-filters are designed to pair with each comb-filter. The HSR of the k-th channel in subband s, denoted by $q_{s,t}(k)$ is computed using Eq. (7). For a comb-function whose inter-peak frequency is close to the input signal's true pitch value, its HSR would be high.

$$q_{s,t}(k) = \sum_f \left|X_{k,s,t}(f)c_k(f)\right|^2 / \sum_f \left|X_{k,s,t}(f)(1-c_k(f))\right|^2 \tag{7}$$

where $X_{k,s,t}(f)$ and $c_k(f)$ denote the DFT coefficients in frame t for input speech and comb-filter respectively.

In the presence of noise, one frame of speech might be more corrupted than its neighboring frames, such that inter-frame peak consistency in $q_{s,t}(k)$ is affected. Since the pitch contour of natural speech generally varies smoothly in time, we applied a lowpass IIR filter on $q_{s,t}(k)$ to improve its inter-frame peak consistency.

Channel selection is performed on a per stream basis, using a three stages selection process designed to improve both pitch estimation and voicing detection performance. The channel is selected according to $q_{s,t}(k)$ and autocorrelation (ACR). The detail of channel selection is introduced in [20].

After channel selection, an HSR-based weighted averaging scheme is performed. Through this weighting scheme, ACRs from the more reliable channels will have a greater impact on their stream SC, resulting in a more prominent ACR peak at the most likely pitch period of the signal. Peak prominence in each stream SC is also due to the use of harmonically-enhanced comb-filtered signals in the individual channel's ACR computation.

To improve the noise-robustness of V/UV detection and pitch estimation, we select the subband which has the high voicing strength. The voicing strength is confirmed based on selected channel. A constant threshold is applied on the maximum interpolated peak amplitude to obtain the V/UV decision for corresponding subband. The subband selection is implemented for three subbands (1–2 k, 2–3 k and 3–4 k). The first subband (0–1 k) is more robust for pitch detection and is always selected. It is effective to eliminate interfere of periodicity noise.

The stream SCs are further fused into a single SC. The contribution of stream SCs are determined by a stream-reliability-weighting function. It is performed in selected subband. It is observed that the maximum HSRs in the more reliable streams are higher. In addition, reliable ACR tend to have similar peak location. Therefore, the weight is made dependent on two factors which include $q_{s,t}(k)$ and autocorrelation (ACR). The stream-reliability-weighting scheme reduces the variability of the maximum peak amplitude in the MBSC for noisy speech, such that this amplitude becomes a robust indicator of the frame's degree of voicing. To improve the inter-frame consistency of the MBSC's maximum peak location across a continuous voiced segment, the same lowpass IIR filter is applied ACR. Time-smoothing also slows down the rate of decrease of peak amplitudes, which in turn improves the detection of weak voiced frames at voicing offsets.

The pitch candidates corresponding to the 10 highest peaks are identified. Each peak and its immediate neighbors are fitted by a parabola, and the amplitude and lag position corresponding to the maximum point of this parabola are the refined pitch measurements for the respective pitch candidate. As for V/UV detection, a constant threshold is applied on the maximum interpolated peak amplitude to obtain the initial

V/UV decision for each frame. This is followed by a 5-point median filtering in time on these initial decisions to get the final V/UV detection.

2.3 LSFs Enhancement

The flowchart of LSFs enhancement is shown in Fig. 3.

The idea behind denoising autoencoders is simple. In order to force the hidden layer to discover more robust features and prevent it from simply learning the identity, we train the autoencoder to reconstruct the input from a corrupted version of it. The denoising auto-encoder is a stochastic version of the auto-encoder. Intuitively, a denoising auto-encoder does two things: try to encode the input and try to undo the effect of a corruption process stochastically applied to the input of the auto-encoder. The latter can only be done by capturing the statistical dependencies between the inputs. The stochastic corruption process randomly sets some of the inputs to zero. Hence the denoising auto-encoder is trying to predict the corrupted values from the uncorrupted values, for randomly selected subsets of missing patterns.

In this subsection, the DAE is applied to LSFs enhancement. There are some advantages for denoising autoencoder. First, by forcing the output to match the original undistorted input data the model can avoid learning the trivial identity solution. Second, since the noises are added randomly, the model learned would be robust to the same kind of distortions in the test data. Third, since each distorted input sample is different, it greatly increases the training set size and thus can alleviate the over fitting problem.

Assuming that $X^{'}$ and $Y^{'}$ are the normalized LSFs of the clean speech and pre-cleaned speech respectively for each frames. The LSFs are normalized according to mean and variable for each dimension. We can formulate the LSFs enhancement process as

$$X^{'} = \eta\left(Y^{'}\right) \tag{8}$$

where η is the LSFs enhancement process that reconstructs the LSFs based on DAE. The objective of LSFs enhancement can be expressed as:

$$\varphi = \text{argmin} E_x\left[\left|\left|\left|\eta\left(Y^{'}\right) - X^{'}\right|\right|_2^2\right] \tag{9}$$

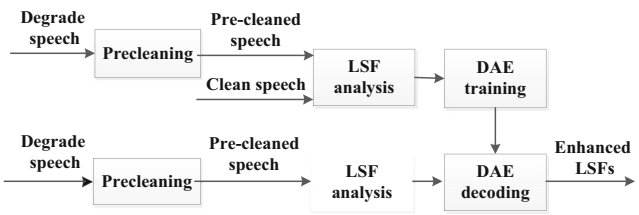

Figure 3 Block diagram of the proposed LSFs enhancement.

The task is to find the φ that is the best estimation of η. We use the speech pair (pre-cleaned LSFs and clean LSFs) to train the DAE. For each hidden layer neural autoencoder, it includes one nonlinear encoding stage and one linear decoding stage as:

$$\begin{aligned} h(y_i) &= \sigma(W_1 y_i + b) \\ x_i &= W_2 h(y_i) + c \end{aligned} \tag{10}$$

where W_1 and W_2 are encoding and decoding matrix as the neural network connection weights. $W_1 = W_2^T$ is used as one regulation. b and c are the vectors of biases of input and output layer, respectively. $\sigma(x_i) = (1 + exp(-x_i))^{-1}$ is the sigmoid activation function and h is the activation of hidden layer.

In the proposed algorithm, a multi hidden layers autoencoder will be trained with the pre-cleaned LSFs as input and the clean LSFs as output. We adopt greedy layer wised pre-training plus fine tuning to train the DAE. The training pair for the first DAE is $X^{'}$ and $Y^{'}$. Then the training pair for the next DAE will be $h(X^{'})$ and $h(Y^{'})$. After pre-training of each layer, all the layers are stacked to form a deep autoencoder for fine tuning. In the fine tuning stage, the initial network parameters are fixed as the parameter obtained from pre-training stage. The parameter is adjusted based on back propagation used in neural network. The DAE is effective to restore the distorted spectrum and enhance LSFs due to the deep models have strong generalization abilities than the shallow models.

2.4 Gain Estimation

The input speech signal gain is measured twice per frame using a pitch adaptive window length [22]. This length is identical for both gain measurements and is determined as follows. For voiced frame, the window length is the shortest multiple of estimated pitch which is longer than 120 samples. For unvoiced frame the window length is 120 samples. The gain calculation for the first window produces and is centered 40 samples before the last sample in the current frame. The calculation for the second window produces and is centered on the last sample in the current frame. The gain is the RMS value, measured in dB, of the signal in the window, s_n.

$$G_i = 10\log_{10}\left(0.01 + \frac{1}{L}\sum_{n=1}^{L} s_n^2\right) \tag{11}$$

where L is the window length.

2.5 Speech Synthesis Based on Multi-band Vocoder

The mixed excitation is implemented using a multi-band mixing model. The effect of this mixed excitation is to reduce the buzz usually associated with LPC vocoder. The relative pulse and noise power in each frequency band is determined

by an estimate of the voicing strength at that frequency in the input speech. The voicing strength for different subband is confirmed according to proposed pitch estimation algorithm in this paper. The mixed excitation is generated as the sum of the filtered pulse and noise excitations. The pulse filter for the current frame is given by the sum of all the bandpass filter coefficients for the voiced frequency bands, while the noise filter is given by the sum of the bandpass filter coefficients for the unvoiced bands.

The adaptive spectral enhancement filter is applied to the mixed excitation signal. This filter is a tenth order pole/zero filters with additional first-order tilt compensation. Its coefficients are generated by bandwidth expansion of the linear prediction filter transfer function corresponding to the interpolated LSF's. Since the excitation is generated at an arbitrary level, the speech gain must be introduced to the synthesized speech. The correct scaling factor is computed for each synthesized pitch period of length. The pulse dispersion filter is a 65th order FIR filter derived from a spectrally flattened triangle pulse. It could reduce the harsh ingredients of synthesized speech.

3 Experiments and Result Analysis

3.1 Data and Analysis Methodology

In this section, we evaluate the proposed approach on pitch estimation and V/UV decision, LSFs enhancement, speech enhancement based on analysis-synthesis framework, parametric speech coding at low bit rate. In this test, the clean speech samples are selected from TIMIT database [23]. Four types of noise recordings extracted from the Noisex-92 database [24], namely pink, factory, volvo and buccaneer, were used as the noise signals. Three SNR conditions, 0 dB, 5 dB and 10 dB, are included in the training process. The speech signal is down-sampled to 8KHz.

For speech enhancement, the frame length is 160 samples and the frame shift is 80 samples. For speech coding, the frame length is the same as mixed excited linear prediction (MELP) standard. The 3000 utterances selected randomly from the training set of the TIMIT database were added with the above mentioned four types of noise and three levels of SNR. Another 300 randomly selected utterances from the TIMIT database were used to construct the test set for each combination of noise types and SNR levels. In addition, we select another 500 randomly selected utterances from the TIMIT database to confirm both the DAE parameter and GMM parameter. Two other noise types, namely white and babble were used for mismatch evaluation.

For pitch estimation and V/UV decision, three types of error metrics are commonly used. The first is Voicing Decision Error (VDE). The second is F0 value estimation error called the Gross Pitch Error (GPE). The FFE takes both GPE and VDE into consideration. [25]

$$VDE = \frac{N_{V\to U} + N_{U\to V}}{N} * 100\% \tag{12}$$

$$GPE = \frac{N_{F0E}}{N_{VV}} * 100\% \tag{13}$$

$$FFE = \frac{N_{VV}}{N} * GPE + VDE \tag{14}$$

For LSFs enhancement, the reference algorithm is Gaussian Mixture Model (GMM) which is used widely in the LSFs transformation. We evaluate the performance of proposed method with the distance between the reconstructed LSFs and the target LSFs according to (15).

$$d^2(lsf_s, lsf_t) = \sum_{i=1}^{10} \omega_i (lsf_{si} - lsf_{ti})^2 \tag{15}$$

where lsf_s and lsf_t represent the reconstructed LSFs parameters and the target LSFs respectively; lsf_{si} and lsf_{ti} represent the *i-th* feature for reconstructed LSFs and target LSFs respectively.

For speech enhancement, the reference algorithm is logMMSE [21] and subspace [26]. We evaluate the performance of proposed method with the perceptual evaluation of speech quality (PESQ) [27]. The PESQ, which is a mean opinion score, is also used to evaluate the quality of the restored speech. It has better correlation with subjective tests than the other objective measures.

In addition, we evaluate the proposed approach in low bite rate speech coding system. The mixed excitation linear prediction (MELP) is the most mature parametric speech coding method so far. Therefore, we select MELP standard to evaluate the proposed algorithm. PESQ is also used for the performance evaluation at low bit rate speech coding.

3.2 DAE Parameter Determination

In this section, we describe the experiments to choose the optimal parameter for DAE model and GMM.

For DAE model, we search over a range of parameter to confirm the number of hidden units (30, 50, 70 and 90) and the number of hidden layers ranging from 1 to 3. For GMM, we search the number of gaussian distribution (16, 32, 64 and 128). In this study, we set the architecture of a DAE as follows: in encoding stage, the size of input layer is 10, each hidden layer is 50. All the layers are stacked and unrolled to form a deep autoencoder layer sizes are 10-50-50-50-10. In our experiments, a batch size of 100 was used. The number of epoch for each layer of pre-training was 20. And in fine tuning stage, the maximum number of iteration was set to 100. We optimize the learning rate ranging from 0.005 to 0.05 (the step is 0.005). The learning rate was set at 0.02.

Table 1 GPE, VDE and FFE of PEAs (SNR=0 dB).

Noise type	Method	VDE(%)	GPE(%)	FFE(%)
Babble noise	GetF0	34.4565	38.3476	52.5654
	MELP	35.1025	40.0155	55.6076
	MBSC	31.7837	22.9145	42.8177
	Proposed	32.9948	17.2508	40.7008
White noise	GetF0	23.4532	23.5632	35.3424
	MELP	22.2663	22.227	33.8609
	MBSC	21.3473	15.5923	28.6778
	Proposed	23.1456	6.9800	26.2740
Volvo noise	GetF0	20.4353	7.3264	24.5335
	MELP	25.1316	10.1734	31.6025
	MBSC	11.711	3.5175	13.8252
	Proposed	10.7854	2.0795	12.0071

We confirm the number of gaussian distribution is 64 for GMM.

3.3 The Evaluation for Pitch Estimation and V/UV Decision

Three other pitch estimation algorithms are also evaluated. The three reference algorithms include: GetF0 [28], MELP [22], and MBSC proposed in [20] which the subband selection and linear prediction residual are not considered. The reference pitch values have been obtained automatically and thoroughly revised manually in the way described in [29]. The performance of the pitch estimation and V/UV decision for different method and different configure as shown in Tables 1 and 2 respectively. Table 1 evaluate the performance in 0 dB condition and Table 2 tabulates the performance, averaged different SNR (0 dB, 5 dB, 10 dB).

Tables 1 and 2 compare performance of the evaluated algorithms on different types of noise. In general, proposed algorithm gives the lowest GPE and FFE in different types of noise. For wide-band noise which include babble and white noise. The MBSC proposed in [20] is prior to the proposed algorithm slightly in terms of VDE. For narrow-band noise which includes Volvo noise, proposed method has lowest rate in terms of VDE. The advantage of proposed algorithm is due to the subband which has the high voicing strength is selected and the linear prediction residual signal is considered in the process of pitch estimation, which can effectively attenuate the noise especially for narrow-band noise.

Table 2 GPE, VDE and FFE of PEAs averaged different SNR.

Noise type	Method	VDE(%)	GPE(%)	FFE(%)
Babble noise	GetF0	29.4364	26.3463	42.5632
	MELP	30.4676	28.0446	45.8182
	MBSC	26.4846	14.4925	33.9186
	Proposed	26.7041	10.8447	31.9626
White noise	GetF0	19.4524	16.3642	28.7864
	MELP	18.5931	15.2937	27.0854
	MBSC	16.4207	9.6668	21.3271
	Proposed	17.5618	4.7345	19.9083
Volvo noise	GetF0	17.4265	7.3425	20.4353
	MELP	23.4122	9.1026	29.2582
	MBSC	10.8934	3.2069	13.0459
	Proposed	10.1208	1.9618	11.2918

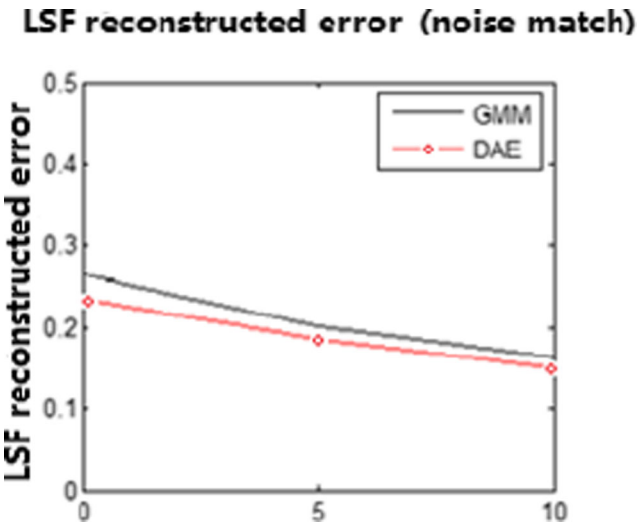

Figure 4 The average LSFs reconstructed error (noise match).

3.4 The Evaluation for LSFs Enhancement

In this subsection, we evaluate the LSFs reconstructed error for different approach which includes GMM and DAE. In this evaluation, different types of noise (pink, factory, buccaneer and volvo for match evaluation; white and babble for

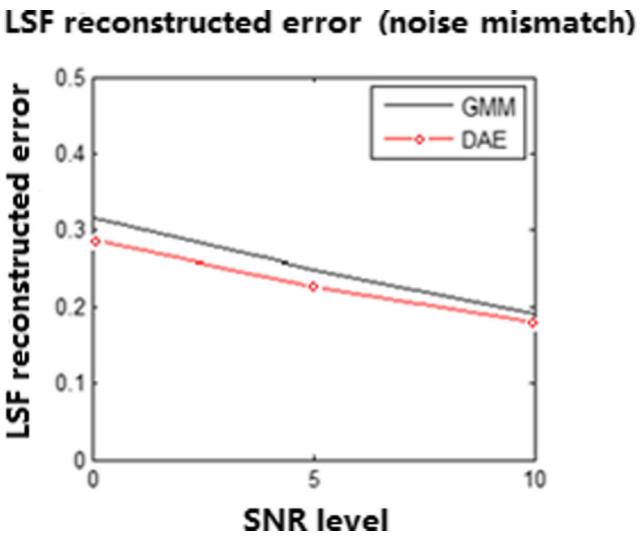

Figure 5 The average LSFs reconstructed error (noise mismatch).

Table 3 The PESQ results for speech enhancement.

Noise type	Method	0 dB	5 dB	10 dB
Average (babble, white factory,volvo)	Noisy	1.7933	2.1925	2.4373
	Logmmse	2.1416	2.5150	2.8619
	Subspace	2.2854	2.5956	2.8943
	Proposed	2.4723	2.6890	2.9817
Pink noise	Noisy	1.6455	2.0673	2.3154
	Logmmse	2.0634	2.4458	2.7807
	Subspace	2.1945	2.5043	2.7584
	Proposed	2.5408	2.7711	2.9131
Buccnaeer noise	Noisy	1.7322	2.1246	2.4544
	Logmmse	2.1928	2.5368	2.8510
	Subspace	2.2542	2.5473	2.8324
	Proposed	2.5164	2.7830	2.9741

The PESQ of match noise is the average PESQ of four type noise which include pink, factory, buccaneer and volvo noise.

mismatch evaluation) are considered. The results of reconstructed error are shown in Figs. 4 and 5. From this figure, we can see that DAE is more effective compared with GMM for both the match noise evaluation and the mismatch noise evaluation. The proposed method is more robust in different noisy environment even not include in training set. The precleaned LSFs could be enhanced effectively through DAE. It is due to the deep models have strong generalization abilities than the shallow models.

3.5 The Evaluation for Speech Enhancement

The proposed method based on analysis-synthesis framework is compared with two different methods, including logMMSE and subspace. The degraded speech without enhancement is denoted as noisy. The PESQ scores are shown in Table 3.

As shown in Table 3, we can see that the proposed method is more effective compared with the different reference methods. It is noted that an average of around 0.3-point improvement over the best conventional method are achieved in various conditions. In comparison with the different reference methods, the proposed method achieves better objective speech quality due to the elimination of musical noise and the restoration of harmonic structure.

3.6 The Evaluation for Low Bit Rate Speech Coding

In this subsection, we applied the proposed algorithm which involve with pitch estimation and LSFs enhancement in low bit rate speech coding. The speech analysis part is different from the MELP standard. The noisy speech is precleaning based on logMMSE firstly; the pitch and V/UV decision is confirmed according to improved MBSC algorithm proposed in this paper. The LSFs is enhanced through DAE model proposed in this paper before vector quantization. The parameter quantization and speech synthesis is the same as MELP standard. The PESQ scores are shown in Table 4.

It is noted that the PESQ MOS score is higher with proposed method in various conditions. In comparison with the MELP standard, the proposed method could achieve better

Table 4 The PESQ results for MELP-2400.

Noise type	Method	0 dB	5 dB	10 dB
Average (babble, white factory,volvo)	Original	1.6934	2.0945	2.3343
	Precleaning	1.9865	2.3665	2.6833
	Proposed	2.3526	2.6615	2.8795
Pink noise	Original	1.5435	1.9653	2.2153
	Precleaning	1.8594	2.2391	2.5397
	Proposed	2.4220	2.6399	2.7794
Buccnaeer noise	Original	1.6342	2.0243	2.3544
	Precleaning	1.9443	2.2964	2.5810
	Proposed	2.3016	2.5628	2.7473

The PESQ of match noise is the average PESQ of four type noise which include pink, factory, buccaneer and volvo noise.

objective speech quality due to the improvement of pitch estimation and the LSFs enhancement.

4 Conclusion and Future Work

In this paper, we present a speech enhancement approach based on analysis–synthesis framework. The proposed approach takes advantage of the analysis–synthesis framework to effectively eliminate musical noise. It builds the mapping relationship to obtain enhanced spectral envelope through deep layer network. The noise could be suppressed and the distorted speech could be restored effectively. The different evaluation results demonstrate the effectiveness of the proposed approach over conventional approaches in various noisy conditions.

In the future, we will improve the current speech enhancement system and focus on noise adaptation in real environment. In addition, we will consider hierarchical neutral network structure according to prior knowledge. We also will consider improving the vocoder structure and expand our algorithm to wideband speech.

Acknowledgments This work is supported by the National High-Tech Research and Development Program of China(863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No.61305003, No.61332017, No.61375027, No.61273288, No.61233009, No.61203258), the Major Program for the National Social Science Fund of China (13&ZD189) and the Integration and application of basic science data in Chinese information processing field (XXH12504-1-11).

References

1. Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing, 27*(2), 113–120.
2. Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing, 32*(6), 1109–1121.
3. Paliwal, K., Schwerin, B., & Wójcicki, K. (2012). Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Communication, 54*(2), 282–305.
4. Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and super gaussian priors. *IEEE Transactions on Speech and Audio Processing, 13*(5), 845–856.
5. Cohen, I. (2002). Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Processing Letters, 9*(4), 113–116.
6. Sreenivas, T. V., & Kirnapure, P. (1996). Codebook constrained Wiener filtering for speech enhancement. *IEEE Transactions on Speech and Audio Processing, 4*(5), 383–389.
7. Mohammadiha, N., Martin, R., & Leijon, A. (2013). Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors. *IEEE Signal Processing Letters, 20*(3), 253–256.
8. Xie, F., & Compernolle, D. V. (1994). A family of MLP based nonlinear spectral estimators for noise reduction. In *Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 53–56). Australia.
9. Dahl, G. E., Sainath, T. N., Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 8609–8613). Canada.
10. Chen, R. F., Chan, C. F., So H. C. (2010). Noise suppression based on an analysis–synthesis approach. In *Proc. Eur. Signal Process. Conf. (EUSIPCO)* (pp. 1539–1543).
11. Cheveigne, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917–1930.
12. Camacho, A., & Harris, J. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *Journal of the Acoustical Society of America, 124*, 1638–1652.
13. Rouat, J., Liu, Y., & Morissette, D. (1997). A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication, 21*, 191–207.
14. Tan, L. N., Alwan, A. (2011). Noise-robust F0 estimation using SNRweighted summary correlograms from multi-band comb filters. In *Proc. IEEE ICASSP* (pp. 4464–4467).
15. Chen, R. F., Chan, C. F., & So, H. C. (2012). Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking. *IEEE Transactions on Speech and Audio Processing, 20*(4), 1324–1336.
16. Toda T., Saruwatari, H., Shikano, K. (2001). Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *Proc of ICASSP* (pp. 941–944).
17. Park, K. Y., & Kim, H. S. (2000). Narrowband to wideband conversion of speech using GMM based transformation. *Proceeding of IEEE International Conference on Acoustics, Speech, Signal Processing, 4*, 1843–1846.
18. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554.
19. Bengio, Y., Yao, L., Alain, G., et al. (2013). Generalized denoising autoencoders as generative models. In *Advances in Neural Information Processing Systems* (pp: 899–907). USA.
20. Tan, L. N., Alwan, A. (2013). Multi-band summary correlogram-based pitch detection for noisy speech. In *Speech communication* (pp. 841–856).
21. Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Speech and Audio Processing, 33*(2), 443–445.
22. Supplee, L. M., Cohn, R. P., Collura, J. S., McCree, A. V. (1997). MELP: the new federal standard at 2400bps. In *Acoustics Speech and Signal Processing* (1591–1594). Germany.
23. Garofolo, J. S. (1993). TIMIT: Acoustic-phonetic Continuous Speech Corpus, Linguistic Data Consortium.
24. Rice University, NOISEX-92 Database, [Online] Available: http://spib.rice.edu/spib/select noise.html.
25. Chu, W., Alwan, A. (2009). reducing F0 frame error of F0 tracking algorithm under noisy condition with an unvoiced/voiced classification frontend. In *Acoustics Speech and Signal Processing* (3969-3972). Germany
26. Jabloun, F., & Champagne, B. (2003). Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing, 11*(6), 700–708.
27. Rix, A. W., Beerends, J. G., Hollier, M. P., et al. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 749–752). USA.
28. Talkin, D. (1995). *Speech Coding and Synthesis*. Elsevier (pp. 497–518).
29. Kotnik, B., Hoge, H., Kacic, Z. (2006). Evaluation of Pitch Detection Algorithms in Adverse Conditions. Proc. 3rd International Conference on Speech Prosody (pp. 149–152). Dresden, Germany.

Bin Liu received his the B.S. degree and the M.S. degree from Beijing institute of technology (BIT), Beijing, in 2007 and 2009 respectively. He received his the Doctor degree from Chinese Academy of Sciences (CAS), Beijing, in 2015. He is currently Research Assistant in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He current research interests include low bit rate speech coding and single channel speech enhancement.

Jianhua Tao received his PhD from Tsinghua University in 2001, and got his Ms from Nanjing University in 1996. He is currently a Professor in NLPR, Institute of Automation, Chinese Academy of Sciences. His current research interests include speech synthesis and coding methods, human computer interaction, multimedia information processing and pattern recognition. He has published more than eighty papers on major journals and proceedings including IEEE Trans. on ASLP, and got several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He serves as the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, NCMMSC etc. He also serves as the steering committee member for IEEE Transactions on Affective Computing, associate editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions, Deputy Editor-in-chief for Chinese Journal of Phonetics.

Zhengqi Wen received his the B.S. degree from University Of Science and Technology of China (USTC), Heifei, in 2008 and received his the Doctor degree from Chinese Academy of Sciences (CAS), Beijing, in 2013. He is currently Research Assistant in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He current research interests include speech recognition and speech synthesis.

Fuyuan Mo received his the B.S. degree from Nanjing University, Nanjing, in 1964. He is a Professor in Institute of Acoustic, Chinese Academy of Sciences, Beijing. He current research interests include speech coding at very low bit rate and underwater communication technology.