Model-Based Speech Enhancement with Improved Spectral Envelope Estimation via Dynamics Tracking

Ruofei Chen, *Student Member, IEEE,* Cheung-Fat Chan, *Member, IEEE* and Hing Cheung So, *Senior Member, IEEE*

***Abstract*—In this work, we present a model-based approach to enhance noisy speech using an analysis-synthesis framework. Target speech is reconstructed with model parameters estimated from noisy observations. In particular, spectral envelope is estimated by tracking its temporal trajectories in order to improve the noise-distorted short-time spectral amplitude. Initially, we propose an analysis-synthesis framework for speech enhancement based on harmonic noise model (HNM). Acoustic parameters such as pitch, spectral envelope, and spectral gain are extracted from HNM analysis. Spectral envelope estimation is improved by tracking its line spectrum frequency trajectories through Kalman filtering. System identification of Kalman filter is achieved via a combined design of codebook mapping scheme and maximum likelihood estimator with parallel training data. Complete system design and experimental validations are given in details. Through performance evaluation based on a study of spectrogram, objective measures and a subjective listening test, it is demonstrated that the proposed approach achieves significant improvement over conventional methods in various conditions. A distinct advantage of the proposed method is that it successfully tackles the "musical tones" problem.**

***Index Terms*—Harmonic noise model, Speech analysis, Speech synthesis, Kalman filter, Codebook mapping, Vector quantization (VQ).**

I. Introduction

SPEECH enhancement is concerned with improving the quality and intelligibility of speech degraded in the presence of background noise. It has been widely studied for decades and various algorithms have been proposed. Among them, the short-time spectral amplitude (STSA) based methods attract a great deal of interest [1][2][3] and generally outperform algorithms in other categories in various noisy conditions [4][5]. The main reason is twofold. First, the former are optimized in a best spectral magnitude estimator sense by noticing the unimportance of the phase in speech enhancement [6]. Second, they take advantage of *a priori* knowledge estimated using a Bayesian framework. However, due to inaccurate noise and *a priori* signal-to-noise ratio (SNR) estimation, STSA-based algorithms often suffer poor performance in non-stationary and/or low SNR environments. It is observed in STSA-based approaches that there is always a trade-off between noise suppression and speech naturalness. The original STSA and log-spectral amplitude (LSA) estimator by Ephraim and Malah [1][2] preserve relatively high level of speech naturalness as they are able to maintain relatively more formant and harmonic information. However, the side effect is that many annoying artifacts (known as "musical tones") and residual noise are also remained. In contrast, subsequent STSA variants that incorporating speech presence uncertainty (SPU)[3][7] are shown to be superior in terms of noise removal and musical tone elimination capability. However, the price to pay is that they further distort the harmonic structure as well as the spectral envelope, which would potentially penalize the signal quality and intelligibility. This may also account for the reason why the target speech processed by these approaches often sounds clean but unnatural.

In recent years, attempts have been made to incorporate harmonic structure of speech in speech enhancement [8][9]. These methods generally work as a post-processing tool that is combined with classical *a priori* SNR estimation to restore missing harmonics that are deteriorated both by noise and by conventional enhancement process. In [10][11], a complete analysis-synthesis framework based on harmonic noise model (HNM) is proposed to re-synthesize clean speech signals based on acoustic cues (e.g. pitch, spectral gain and spectral envelope) extracted from noisy observations. In doing so, target speech is reconstructed with speech related information only and background noise is automatically removed. This approach is attractive as it can retrieve damaged harmonic structure and at the same time eliminates residual noises and musical tones. However, to ensure accurate model parameter estimation, a pre-processing filter is often required to pre-clean the noisy signals prior to HNM analysis. It is reported in [10][11] that pitch and spectral gain estimation applied on pre-cleaned spectrum can give satisfactory result even in very low SNR environments. Nevertheless, most pre-cleaning algorithms (conventional speech enhancement methods) fail to recover the spectral envelope which has been distorted by background noise. In even worse conditions, noise-corrupted spectral envelope is further modified by the pre-cleaning process, results in mismatched harmonic magnitude generation. As a consequence, improved spectral envelope estimation is desired to restore the original spectral shape, so as to improve the overall quality and intelligibility of synthetic speech.

The method to improve the spectral envelope estimation can be regarded as a problem of estimating clean autoregressive (AR) parameters of speech from noisy observations. It is well-known that Wiener filtering is correlated with linear predic-

The authors are with the Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. Email: (ruofechen2@student.cityu.edu.hk; itcfchan@cityu.edu.hk; hcso@ee.cityu.edu.hk).

tion, and clean AR parameters can be iteratively estimated from noisy speech using Wiener filtering [12]. To improve the estimates of clean speech and noise statistics, existing works [13][14] choose to train speech and noise codebooks of AR parameters, which are served as intra-frame constraints for iterative Wiener filtering (IWF). Besides, exploitation of Kalman filter in speech enhancement is also widely studied [15][16][17]. In these methods, speech and/or noise are modeled as stochastic AR process, and AR parameters are represented in state-space form to model the state transition between time samples. Parameter estimation and iterative update can be achieved by expectation-maximization (EM) type algorithms [17]. Their common feature is that they assume constant linear dependence between time domain speech samples within an analysis frame, and make use of Kalman filter to track the intra-frame correlation between speech and noise samples. An attempt has been made in [18] to look at the inter-frame correlation between speech dynamics. In their work, the trajectory of each frequency component is modeled using an AR model, and a Kalman filter is incorporated to track the temporal discrete Fourier transform (DFT) trajectories. Experimental results demonstrate the effectiveness of utilization of inter-frame correlations between speech dynamics. However, employing a Kalman filter for each frequency component in each frequency channel is computationally expensive.

In this paper, we present a speech dynamics tracking approach used in conjunction with HNM based analysis-synthesis framework to enhance noisy speech signals. HNM removes background noise by generating clean harmonics while dynamics tracking scheme further enhances the spectral envelope after noise removal. More specifically, it incorporates Kalman filter to track the temporal trajectories of line spectrum frequencies (LSFs) obtained from linear prediction analysis. The major difference between the proposed method and conventional Kalman tracking approaches is twofold. First, it captures the inter-frame evolution of spectral shapes rather than the intra-frame evolution of time samples. Second, instead of using AR modeling, the linear dependence between state transition and state-observation mapping are modeled by full matrices, respectively. The proposed design is supported by previous investigations on long-term correlations between LSF coefficients [19][20] and by experimental validations, which is discussed in details in Section III. The system identification of Kalman filter is achieved via codebook and maximum likelihood (ML) based offline training. The enhanced LSFs are then directed into the analysis-synthesis framework to improve the spectral envelope estimation, and hence the performance of HNM based speech enhancement.

The distinguishable difference between this proposed design and conventional speech enhancement methods is twofold. On one hand, it takes advantage of the analysis-synthesis framework to effectively eliminate musical tones and residual noise. On the other hand, it looks for long-term speech evolution to obtain smoothed estimates of spectral shapes through dynamics tracking. Both objective and subjective evaluation results demonstrate the effectiveness of the proposed method over conventional methods in various noisy conditions.

The remainder of this paper is organized as follows. In Section II, the combined design of analysis-synthesis framework and dynamics tracking is presented. In Section III, experimental validation and practical issues such as parameter choice are discussed. In Section IV, the performance of the proposed method is evaluated and compared with conventional methods. Finally, conclusion is drawn in Section V.

II. System Development

The block diagram of the complete system design is shown in Fig.1. During enhancement, noisy speech is initially pre-cleaned on a frame basis. HNM analysis is then applied and acoustic cues such as pitch, spectral envelope (harmonic magnitude), and spectral gain are estimated from the pre-cleaned signals. To further improve the spectral distortion introduced by additive noise as well as the pre-cleaning process, a dynamic tracking scheme is incorporated. For each frame, it looks for a block of LSFs observed up to this frame to form the feature matrix of noisy observation. A link is established between this observation and its matched Kalman filter parameters via offline training. Online Kalman adaptation is applied and smooth estimates of enhanced LSFs of current frame are obtained. Spectral envelope constructed by the enhanced LSFs is adopted to replace the original one which is estimated from pre-cleaned signals. All estimated and enhanced speech parameters are sent to the synthesizer to reconstruct the target speech. Detailed procedures for both the HNM-based framework and the dynamic tracking scheme are shown as follows.

A. HNM based Analysis-Synthesis Framework

HNM is a speech production model that is widely exploited in speech coding and synthesis. In these applications, with loose bit-rate requirement, very high quality of speech can be reproduced with HNM modeling, owing to its flexible and effective decomposition of speech. It assumes the speech signal to be composed of a deterministic/voiced part and of a stochastic/unvoiced part. The voiced part is assumed to contain only harmonically related sinusoids while the unvoiced part can be modeled by random signals [21]. The major motivation of applying HNM in speech enhancement is to take advantage of its organized harmonic structure as *a priori* knowledge to improve the estimate of clean speech signals degraded by additive noise. The proposed speech enhancement framework comprises two stages, namely, speech analysis and speech synthesis. At speech analysis stage, the general idea is to extract only the speech related features from noisy observations and send them to the HNM synthesizer. In practice, the acoustic parameters of interest are pitch, harmonic magnitude, spectral gain and voiced/unvoiced (V/UV) information. It is worth mentioning that, in practice, it is very difficult to directly apply HNM analysis in adverse environment (e.g., SNR $\leq$ 10dB). In such cases, a preliminary de-noising step is required to pre-clean the noisy signals. The major steps of the proposed HNM-based speech enhancement system are given as follows.

1) Pre-cleaning: The goal of the pre-cleaning step is to filter the noisy signals such that it is more suitable for the HNM analysis. In this sense, there are two major reasons for

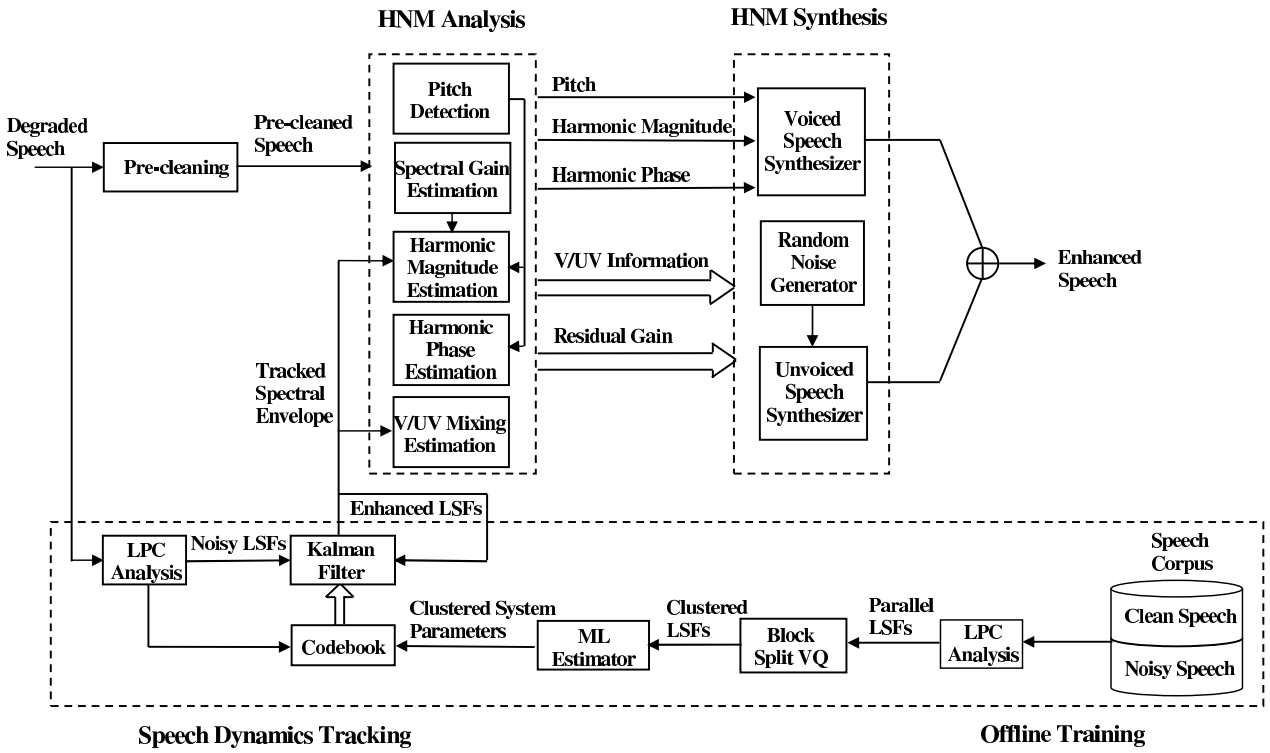

Fig. 1: Block diagram of the proposed system design

incorporating a pre-processing filter in the proposed design. First, the pre-cleaning step de-emphasizes the portion of spectrum which is dominated by noise for frequency domain pitch detection algorithms. In doing so, it improves the accuracy of spectrum matching error function in noisy conditions and thereby contributes to robust pitch estimation. Second, it provides a rough estimation of speech and noise statistics and therefore it is possible to estimate an overall spectral gain which is close to that of clean speech. Experimental results in [10][11] show that STSA-based methods are basically capable of doing the work.

2) Pitch detection: Due to the inaccuracy in noise and SNR estimation, it is typical that in the STSA enhanced magnitude spectra, only harmonic bands with high *a priori* SNR are retained while the rest of spectrum are flattened. Fig.2 illustrates this phenomenon by showing the short-time log-magnitude spectra of a voiced speech segment, and its degraded (by 0dB white noise) as well as pre-cleaned (by [3]) version. Motivated by the fact that several dominant harmonics and their frequency locations are preserved after pre-cleaning, it is able to develop a frequency domain pitch detection algorithm based on [22] to match certain portion of the pre-cleaned spectrum with the excitation spectrum. The optimum pitch period τ_0 is obtained by minimizing an error function $\alpha(\tau)$ with respect to the searching variable τ as

$$\tau_0 = \arg\min_{\tau}\{\alpha(\tau)\} \tag{1}$$

The cost function is constructed by matching the input speech spectrum $S(k)$ with the pitch-dependent synthetic excitation

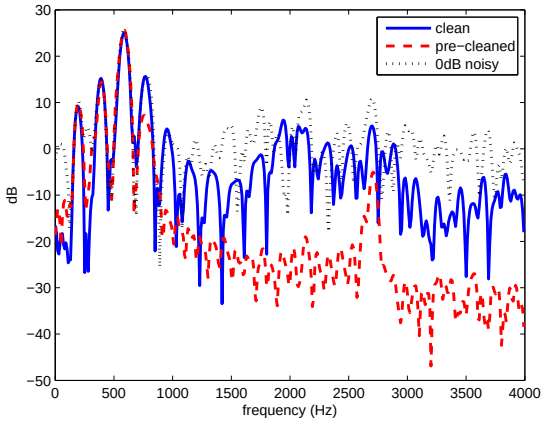

Fig. 2: Short-time log-magnitude spectra of original clean, noisy (white noise, SNR = 0dB), and pre-cleaned voiced speech signals

spectrum $E(\tau, k)$ as

$$\alpha(\tau) = \frac{\sum_{m=1}^{M(\tau)} w_m(\tau) \sum_{k=a_m(\tau)}^{b_m(\tau)} [|S(k)| - A_m(\tau)|E(\tau,k)|]^2}{(1-\tau B)\sum_{m=1}^{M(\tau)} \sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2} \tag{2}$$

where $M(\tau)$ is the total number of harmonic bands, $a_m(\tau)$ and $b_m(\tau)$ are the lower and upper boundaries of the m^{th} harmonic

band, B is a weighting factor for biasing the pitch dependent error, $w_m(\tau)$ is a frequency dependent weight that is imposed to selectively emphasize on various frequency regions, and $A_m(\tau)$ is the harmonic magnitude obtained by minimizing the matching error in each harmonic band, which gives:

$$A_m(\tau) = \frac{\sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)||E(\tau,k)|}{\sum_{k=a_m(\tau)}^{b_m(\tau)} |E(\tau,k)|^2} \tag{3}$$

By noticing the characteristics of pre-cleaned spectra, $w_m(\tau)$ can be derived with $a\,priori$ SNR estimated in each harmonic band. In doing so, the retained harmonic region would contribute more to the overall error function as compared to remaining over-suppressed regions. However, as pointed out in [23], (2) does not penalize the mismatch between input and excitation for small energy harmonic bands, in case they are located between two adjacent high energy voiced bands. As a consequence, gross pitch errors such as double pitch errors may occur. To tackle this issue, a similar corrective error function $\beta(\tau)$ is also required to emphasize the mismatch by normalizing the band energy before applying weights:

$$\beta(\tau) = \gamma(\tau) \sum_{m=1}^{M(\tau)} w_m(\tau) \left[\frac{\sum_{k=a_m(\tau)}^{b_m(\tau)} [|S(k)| - A_m(\tau)|E(\tau,k)|]^2}{\sum_{k=a_m(\tau)}^{b_m(\tau)} |S(k)|^2} \right] \tag{4}$$

where $\gamma(\tau) = [M(\tau)(1-\tau B)]^{-1}$. There is always a trade-off between the emphasis on voiced and unvoiced bands, and a compromise is reached by combining these two measures, resulting in the final error function $\epsilon(\tau)$:

$$\epsilon(\tau) = \lambda\alpha(\tau) + (1-\lambda)\beta(\tau) \tag{5}$$

where $\lambda \in (0,1)$ is a weighting factor.

3) Harmonic magnitude estimation: In clean conditions, given the optimum pitch period τ_0, harmonic magnitudes could be estimated straightforwardly using (3). However, in adverse conditions, the use of either noisy or pre-cleaned spectrum would result in large matching errors, and hence significantly degrades the enhancement performance. To tackle this issue, magnitude spectra are modeled using linear predictive coding (LPC) spectral envelopes in this work. In doing so, spectral modification can be achieved by simply manipulating the LPC coefficients. This configuration allows us to incorporate a time-frequency tracking scheme to re-estimate the envelope spectrum with reduced dimensions. Consequently, in the proposed design, harmonic magnitude for the each harmonic band is determined by sampling the enhanced LPC envelope spectrum at each integer multiples of pitch frequencies. While the enhanced LPC envelope spectrum is obtained by tracking its LSF trajectories and the detailed procedure of the dynamic tracking scheme is discussed in Section II-B.

4) Spectral gain estimation: In LPC model, over an analysis frame, speech signals are modeled as a combination of vocal tract parameters (spectral envelope) and an excitation gain. The excitation gain indicates the overall energy level of current frame while the energy-normalized spectral envelope represents the spectral shape. By taking advantage of this decomposition, spectral envelope and excitation gain can be adjusted independently within the proposed analysis-synthesis framework. Assume that $S_\ell(m)$ are the sampled input magnitude spectrum and $\bar{S}_\ell(m)$ are the sampled energy-normalized envelope spectrum, both having a total of $M_\ell(\tau_0)$ harmonic bands at the ℓ^{th} frame, respectively. The overall gain g_ℓ is obtained by minimizing

$$\sum_{m=1}^{M_\ell(\tau_0)} [|S_\ell(m)| - g_\ell|\bar{S}_\ell(m)|]^2 \tag{6}$$

with respect to g_ℓ, giving

$$g_\ell = \left(\sum_{m=1}^{M_\ell(\tau_0)} |S_\ell(m)||\bar{S}_\ell(m)| \right) \left(\sum_{m=1}^{M_\ell(\tau_0)} |\bar{S}_\ell(m)|^2 \right)^{-1} \tag{7}$$

5) V/UV mixing function: HNM is able to remove background noise and then generate clean harmonics. However, due to the physical mechanism of human speech production, noise-like components such as fricatives also occur in voiced utterance. Therefore, pure clean harmonic generation would introduce buzziness in synthetic speech. To tackle this issue, controllable amount of noise could be artificially inserted to compensate this effect. To achieve this, a V/UV mixing function within each frame is imposed to allow certain amount of noise to be added in the harmonic portion of speech spectrum. It has been shown in [24] that clean speech spectral envelope inherently correlates to the degree of V/UV mixture and a spectral flatness measure is defined as

$$f(\theta) = \frac{1}{\pi - \theta} \int_\theta^\pi [\log|\bar{S}(\omega)| - m(\theta)]^2 d\omega \tag{8}$$

where

$$m(\theta) = \frac{1}{\pi - \theta} \int_\theta^\pi \log|\bar{S}(\omega)| d\omega \tag{9}$$

By comparing to a predefined threshold T_{uv}, a smooth V/UV mixing function $\rho(\theta)$ is defined from the spectral flatness measure as:

$$\rho(\theta) = \begin{cases} 1 - \dfrac{f(\theta)}{2T_{uv}}, & f(\theta) < T_{uv} \\ \dfrac{T_{uv}}{2f(\theta)}, & f(\theta) > T_{uv} \end{cases} \tag{10}$$

Note that the V/UV transition is at $\rho(\theta) = 0.5$, and the voiced and unvoiced regions are corresponding to $\rho(\theta) < 0.5$ and $\rho(\theta) > 0.5$, respectively. In this proposed design, the mixing function is derived with the envelope spectrum constructed by the tracked LSFs, and small amount of randomly generated white Gaussian noise, which is weighted by this mixing function, is added to the synthetic harmonic spectrum.

6) Voiced speech synthesis: At synthesis stage, voiced and unvoiced signals are reconstructed with different strategies. A time domain approach is adopted as suggested in [22] to allow continuous variation in HNM parameters. Voiced speech signals at time instant t of the ℓ^{th} frame is re-synthesized using a sum of sinusoids running at the harmonics of the estimated pitch frequency as

$$v_\ell(t) = \sum_{m=1}^{M(\tau_0)} \hat{A}_{\ell,m}(t) cos[\hat{\theta}_{\ell,m}(t)] \tag{11}$$

where $\hat{A}_{\ell,m}(t)$ and $\hat{\theta}_{\ell,m}(t)$ are the estimated m^{th} harmonic magnitude and phase, respectively. Harmonic magnitudes at each frame index are sampled from the enhanced envelope spectrum which is constructed from the tracked LSFs with a pre-cleaned excitation gain. Harmonic phases extracted from noisy input signals are employed as phase information is less important in speech enhancement [6]. For each intra-frame time instant t along temporal track, the time-varying magnitude function $\hat{A}_{\ell,m}(t)$ is linearly interpolated while the time-varying phase function $\hat{\theta}_{\ell,m}(t)$ is quadratically interpolated based on linear interpolation of harmonic frequencies.

7) Unvoiced speech synthesis: A frequency domain approach is used for unvoiced speech synthesis. The enhanced envelope spectrum is weighted by the V/UV mixing function and is then converted to autocorrelation data. An all-pole LPC model is fitted to compute the residual gain of the synthesis filter. Random Gaussian noise is generated and fed into the synthesis filter to produce the unvoiced portion of speech. The final target speech is produced by simply summing the voiced and unvoiced speech signals in time domain.

B. Speech Dynamics Tracking

As discussed in the previous section, the highly distorted envelope spectrum is still yet to be improved to accurately estimate the harmonic magnitude for HNM based speech enhancement system. LSFs obtained from linear prediction analysis are widely used in speech applications to represent the spectral envelope. Previous works [19][20] have revealed certain long-term correlations between consecutive LSFs, so it is reasonable to exploit a linear dynamical system (LDS) model to track the temporal LSF trajectories. In addition, ordering and stability properties of LSFs also make it suitable for recursive filtering problem.

1) Kalman tracking: In this LDS modeling, noisy and clean speech signals are chopped into sliding blocks and each block contains K overlapping frames. Both the block shift and the frame shift are T samples. It is assumed that within an analysis block, there exists certain inter-frame (between consecutive clean LSFs i.e., $\mathbf{x}_\ell$ and $\mathbf{x}_{\ell+1}$) and intra-frame (between noisy and clean LSFs i.e., $\mathbf{y}_\ell$ and $\mathbf{x}_\ell$) linear relationships. In this context, the clean LSFs $\mathbf{x}_\ell$ of the ℓ^{th} frame are modeled as the internal states and the noisy LSFs $\mathbf{y}_\ell$ are modeled as the observations, and they can be represented in state-space form as:

System dynamic model:

$$\begin{aligned} \mathbf{x}_{\ell+1} &= \mathbf{F}\mathbf{x}_\ell + \mathbf{w}_\ell \\ \mathbf{w}_\ell &\sim \mathcal{N}(0, \mathbf{Q}) \end{aligned} \tag{12}$$

Measurement model:

$$\begin{aligned} \mathbf{y}_\ell &= \mathbf{H}\mathbf{x}_\ell + \mathbf{v}_\ell \\ \mathbf{v}_\ell &\sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \tag{13}$$

where $\mathbf{F}$ is the state transition matrix and $\mathbf{H}$ is a linear mapping matrix for state and observation. Additionally, it is assumed that $\mathbf{w}_\ell$ and $\mathbf{v}_\ell$ are uncorrelated, zero-mean Gaussian vectors with covariances $\mathbf{Q}$ and $\mathbf{R}$. Thus we can construct our best guess of state $\mathbf{x}_\ell$ and its covariance $\mathbf{\Sigma}_\ell$ at the ℓ^{th} frame based on data observed until $s \leq \ell$, that is

$$\hat{\mathbf{x}}_{\ell|s} = E[\mathbf{x}_\ell|\mathbf{y}_s] \tag{14}$$

$$\mathbf{\Sigma}_{\ell|s} = Cov[\mathbf{x}_\ell|\mathbf{y}_s] = E[(\mathbf{x}_\ell - \hat{\mathbf{x}}_{\ell|s})(\mathbf{x}_\ell - \hat{\mathbf{x}}_{\ell|s})^T|\mathbf{y}_s] \tag{15}$$

Then the transformations between noisy and clean LSFs are characterized by a set of Kalman recursion equations described as:

$$\hat{\mathbf{x}}_{\ell|\ell-1} = \mathbf{F}\hat{\mathbf{x}}_{\ell-1|\ell-1} \tag{16}$$

$$\mathbf{\Sigma}_{\ell|\ell-1} = \mathbf{F}\mathbf{\Sigma}_{\ell-1|\ell-1}\mathbf{F}^T + \mathbf{Q} \tag{17}$$

$$\mathbf{e}_\ell = \mathbf{y}_\ell - \mathbf{H}\hat{\mathbf{x}}_{\ell|\ell-1} \tag{18}$$

$$\mathbf{\Sigma}_{e_\ell} = \mathbf{H}\mathbf{\Sigma}_{\ell|\ell-1}\mathbf{H}^T + \mathbf{R} \tag{19}$$

$$\mathbf{K}_\ell = \mathbf{\Sigma}_{\ell|\ell-1}\mathbf{H}^T\mathbf{\Sigma}_{e_\ell}^{-1} \tag{20}$$

$$\hat{\mathbf{x}}_{\ell|\ell} = \hat{\mathbf{x}}_{\ell|\ell-1} + \mathbf{K}_\ell\mathbf{e}_\ell \tag{21}$$

$$\mathbf{\Sigma}_{\ell|\ell} = \mathbf{\Sigma}_{\ell|\ell-1} - \mathbf{K}_\ell\mathbf{\Sigma}_{e_\ell}\mathbf{K}_\ell^T \tag{22}$$

Given a series of noisy observations for each analysis block, the basic idea is to train a specific set of system parameters $\mathbf{\Theta}$ to initialize the Kalman filter/smoother. The system parameters include $\mathbf{\Theta} = \{\mathbf{F}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \hat{\mathbf{x}}_1, \mathbf{\Sigma}_1\}$, where $\hat{\mathbf{x}}_1$ and $\mathbf{\Sigma}_1$ are initial state mean and error covariance, respectively. Subsequently, with a series of noisy observations and system parameters $\mathbf{\Theta}$ at hand, one performs the above set of Kalman recursion equations to obtain clean estimates of LSFs. In addition, a set of backward recursions [25] could be performed to obtain a smoothed non-causal estimate of clean LSFs.

2) System identification: In this work, an offline learning system is designed to train a codebook to initialize the Kalman filter that runs online adaptation. The proposed design is constructed using the framework of well-known Linde-Buzo-Gray (LBG) algorithm [26] with split vector quantization (VQ). Nevertheless, the major difference is that a block concept is introduced in the proposed design. In conventional LBG algorithm, vectors of LSFs are clustered in a sense that frames with similar spectral patterns are grouped together. Whereas in this design, blocks of vectors of LSFs are clustered so that matched temporal correlations between blocks are also taken into account when calculating the distortion measure. Initially, a parallel database is adopted, both noisy and clean signals are chopped into sliding blocks with overlapping frames according to the proposed Kalman filtering structure. The focus of this design is to capture and cluster the spectral shape evolution that is independent of the impact of overall spectral energy level. Consequently in linear prediction analysis, autocorrelation coefficients of each frame are normalized by its short-time energy and LSFs are computed from the normalized data to achieve energy-independent training. Let a $P \times K$ matrix

$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K]$ denote noisy and clean LSF blocks, respectively, where P is the linear prediction order, and K is the number of frames per block. Intuitively, the distortion measure between training blocks and codebook entries should be constructed by minimizing a sum of total log spectral distance (LSD) within this block. Nevertheless, this implementation is computationally expensive, and also the centroid of LSD measure is difficult to compute. Based on the modified Itakura-Saito (IS) measure [26], an approximate quadratic distortion measure between direct LPC filter form of noisy block $\mathbf{A}$ and the j^{th} entry in codebook $\hat{\mathbf{A}}_j$ is defined as:

$$d(\mathbf{A}, \hat{\mathbf{A}}_j) = \sum_{k=1}^{K} (\mathbf{a}_k - \hat{\mathbf{a}}_{j,k})^T \mathbf{R}_k (\mathbf{a}_k - \hat{\mathbf{a}}_{j,k}) \tag{23}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$, $\hat{\mathbf{A}}_j = [\hat{\mathbf{a}}_{j,1}, \ldots, \hat{\mathbf{a}}_{j,K}]$, and the weighting matrix $\mathbf{R}_k$ is the autocorrelation matrix. Theoretical analysis of LPC parameters in [27] shows that (23) can be reformulated as a quadratic measure between LSFs as

$$d(\mathbf{Y}, \hat{\mathbf{Y}}_j) = \sum_{k=1}^{K} (\mathbf{y}_k - \hat{\mathbf{y}}_{j,k})^T \mathbf{W}_k (\mathbf{y}_k - \hat{\mathbf{y}}_{j,k}) \tag{24}$$

where $\mathbf{W}_k = \mathbf{J}_k^T \mathbf{R}_k \mathbf{J}_k$ is the sensitivity matrix with $\mathbf{J}_k$ being the Jacobian matrix transforming LSFs to direct LPC coefficients [27]. There are two reasons to adopt the LSF form rather than the direct LPC form. First, diagonalized sensitivity matrix indicates scale quantization of LSFs does not affect each other, and hence results in less quantization error. Second, the weighted mean square error (WMSE) is easy to compute compared to general quadratic measure. As a result, an input noisy block $\mathbf{Y}$ is clustered based on a codebook searching index $j_{\min}$, which is defined as

$$j_{\min} = \arg\min_j \{d(\mathbf{Y}, \hat{\mathbf{Y}}_j)\} \tag{25}$$

Assume that a total of L LSF blocks are grouped into the chosen cluster. The block centroid of this specific cluster $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_K]$ is then obtained by sequentially minimizing

$$\sum_{i=1}^{L} (\mathbf{y}_{i,k} - \hat{\mathbf{y}}_k)^T \mathbf{W}_{i,k} (\mathbf{y}_{i,k} - \hat{\mathbf{y}}_k) \tag{26}$$

which results in

$$\hat{\mathbf{y}}_k = (\sum_{i=1}^{L} \mathbf{W}_{i,k})^{-1} (\sum_{i=1}^{L} \mathbf{W}_{i,k} \mathbf{y}_{i,k}) \tag{27}$$

One performs LBG algorithm to iteratively update (24)-(27) and subsequently obtain a parallel training subset of both observation and state LSF blocks for each cluster as $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_L, \mathbf{X}_1, \ldots, \mathbf{X}_L\}$. This indicates that for each sequence of observed noisy LSFs, a specific set of noisy blocks with similar spectral patterns and evolution trajectories as well as their original clean pairs can be identified. Assume the prediction error $\mathbf{e}_\ell$ in (18) has a multivariate normal distribution so that parameter estimation for Kalman filter can be achieved with ML estimator. More specifically, for each cluster, the system parameters $\mathbf{\Theta}$ are obtained by minimizing the total negative log likelihood for all L blocks of observations, which is given by:

$$\begin{aligned}
\mathcal{J}(\mathbf{X}, \mathbf{Y}, \mathbf{\Theta}) &= -\sum_{i=1}^{L} \mathcal{L}(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{\Theta}) \\
&= \sum_{i=1}^{L} \sum_{\ell=2}^{K} (\mathbf{x}_\ell^{(i)} - \mathbf{F}\mathbf{x}_{\ell-1}^{(i)})^T \mathbf{Q}^{-1} (\mathbf{x}_\ell^{(i)} - \mathbf{F}\mathbf{x}_{\ell-1}^{(i)}) \\
&+ \sum_{i=1}^{L} \sum_{\ell=1}^{K} (\mathbf{y}_\ell^{(i)} - \mathbf{H}\mathbf{x}_\ell^{(i)})^T \mathbf{R}^{-1} (\mathbf{y}_\ell^{(i)} - \mathbf{H}\mathbf{x}_\ell^{(i)}) \\
&+ \sum_{i=1}^{L} (\mathbf{x}_1^{(i)} - \hat{\mathbf{x}}_1)^T \mathbf{\Sigma}_1^{-1} (\mathbf{x}_1^{(i)} - \hat{\mathbf{x}}_1) + L \ln |\mathbf{\Sigma}_1| \\
&+ L(N-1) \ln |\mathbf{Q}| + LN \ln |\mathbf{R}| + \text{constant}
\end{aligned}$$

Assume that there is no constraint imposed on the system matrices, the estimates of $\mathbf{\Theta}$ are derived as a multiple-observation extension of the results obtained in [28] as

$$\mathbf{F} = \left(\sum_{i=1}^{L} \sum_{\ell=2}^{K} \mathbf{x}_\ell^{(i)} \mathbf{x}_{\ell-1}^{(i)T} \right) \left(\sum_{i=1}^{L} \sum_{\ell=2}^{K} \mathbf{x}_{\ell-1}^{(i)} \mathbf{x}_{\ell-1}^{(i)T} \right)^{-1} \tag{28}$$

$$\mathbf{H} = \left(\sum_{i=1}^{L} \sum_{\ell=1}^{K} \mathbf{y}_\ell^{(i)} \mathbf{x}_\ell^{(i)T} \right) \left(\sum_{i=1}^{L} \sum_{\ell=1}^{K} \mathbf{x}_\ell^{(i)} \mathbf{x}_\ell^{(i)T} \right)^{-1} \tag{29}$$

$$\mathbf{Q} = \frac{1}{L(K-1)} \left(\sum_{i=1}^{L} \sum_{\ell=2}^{K} \mathbf{x}_\ell^{(i)} \mathbf{x}_\ell^{(i)T} - \mathbf{F} \sum_{i=1}^{L} \sum_{\ell=2}^{K} \mathbf{x}_{\ell-1}^{(i)} \mathbf{x}_\ell^{(i)T} \right) \tag{30}$$

$$\mathbf{R} = \frac{1}{LK} \left(\sum_{i=1}^{L} \sum_{\ell=1}^{K} \mathbf{y}_\ell^{(i)} \mathbf{y}_\ell^{(i)T} - \mathbf{H} \sum_{i=1}^{L} \sum_{\ell=1}^{K} \mathbf{x}_\ell^{(i)} \mathbf{y}_\ell^{(i)T} \right) \tag{31}$$

$$\hat{\mathbf{x}}_1 = \frac{1}{L} \sum_{i=1}^{L} \mathbf{x}_1^{(i)} \tag{32}$$

$$\mathbf{\Sigma}_1 = \frac{1}{L} \sum_{i=1}^{L} \mathbf{x}_1^{(i)} \mathbf{x}_1^{(i)T} - (\hat{\mathbf{x}}_1)(\hat{\mathbf{x}}_1)^T \tag{33}$$

3) Iterative update: To achieve iterative update of the tracking scheme, instead of using (32) as the initial guess, enhanced LSFs of last analysis block are used as the initial state LSFs for the next block.

III. Validation and Practical Issues

A. Robustness of HNM Parameter Estimation

The performance of the proposed speech enhancement system relies heavily on the estimation accuracy of acoustic parameters, and to a large extent, the accuracy of pitch frequency and harmonic magnitude estimation. The proposed pitch detection algorithm offers flexibility in weighting matching errors in each harmonic band. As a result, it can be customized for specific pre-cleaning algorithms, and hence robust pitch detection results can be obtained in different adverse conditions. It is confirmed from our experiment that over 90 percent accuracy is achieved for voiced speech segments corrupted by white noise at 0dB SNR level.

In contrast to the spectral envelope which reflects the variations in different frequency regions, the spectral gain

derived in (7) indicates the overall spectral energy level of current frame. It is mentioned that the spectral envelope could be distorted both by noise and the pre-cleaning process. It is interesting to evaluate the impact of overall spectral gain in noisy and pre-cleaned conditions. Fig.3 shows the overall spectral gain contours derived in various conditions. It is observed that this gain fluctuates significantly if it is directly measured from noisy signals. Nevertheless, it approximates the original clean gain contours after applying pre-cleaning. It implies, by means of this envelope-plus-gain decomposition, it is possible to accurately estimate harmonic magnitude by combining enhanced spectral envelope with a pre-cleaned spectral gain. Experiments have been carried out on four variants of STSA-based methods, including the classical STSA method [1], the LSA method [2], a LSA method that incorporating SPU (LSA_SPU) [7], and an optimal modified LSA estimator (OM_LSA) [3]. and it is found that the the LSA estimator gives best gain estimation results when working as a pre-cleaning tool for gain estimation.

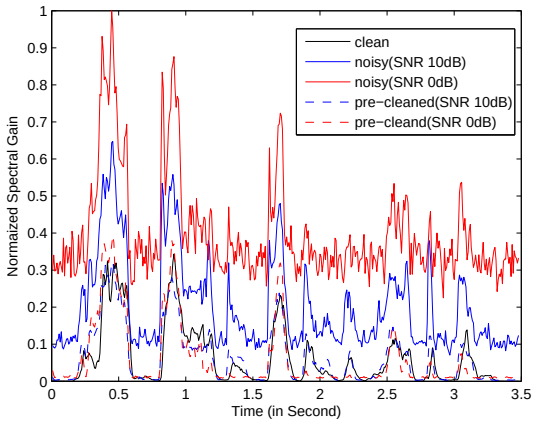

Fig. 3: Spectral gain contours of original clean, noisy (white noise, SNR = 0dB and 10dB), and pre-cleaned speech segments

B. Practical Design of Offline Training

The goal of offline training is to enhance the energy-normalized spectral envelope. The first step is to determine the feature representation of noisy observation for each analysis block. This paragraph explains the reasons why the LSF representation is employed. Within the proposed LDS tracking scheme, it is computationally prohibitive to track the DFT coefficients with sufficient frequency resolution. Alternatively, harmonic magnitudes could be adopted to represent the spectral envelope. However, parameters from LPC analysis are preferred for several reasons. First, the total number of harmonic magnitudes varies from frame to frame, and hence variable length VQ is required. Second, the total number of harmonic magnitudes is pitch dependent, thus gross pitch error could potentially penalize the tracking enhancement. Last but not the least, the order of harmonic magnitude could also make it practically infeasible for speech segments with

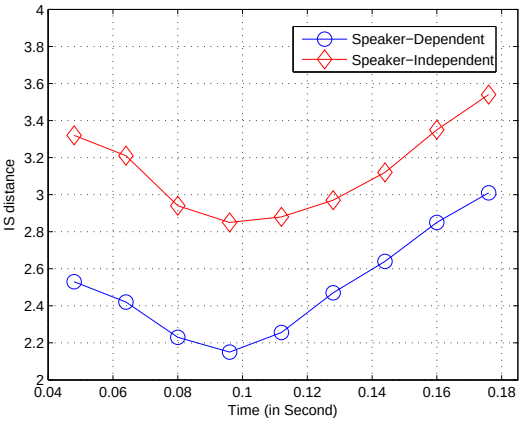

Fig. 4: IS distances for linear dependence test

fairly low pitch (e.g., older male). Conversely, the LSF form of LPC coefficients model the spectral envelope with small dimensions, and the desirable statistical properties of this representation also make it more suitable for quantization and state tracking. It is confirmed in our experiments that LSF coefficients would result in less spectral distortion as compared to direct LPC coefficients and reflection coefficients. Furthermore, the latter could result in unstable tracking of Kalman filters. Consequently, only LSF representation is adopted and evaluated in this validation.

The next step is to determine the effective block length for the proposed design. In the proposed Kalman tracking scheme, constant linear dependence between consecutive clean LSFs within an analysis block is assumed. Therefore, it is desirable to evaluate and determine the best effective block length to validate this assumption. On one hand, short block length is desired such that less modeling and clustering error occurs for the linear state transition. On the other hand, long block length is preferred as more meaningful statistics could be attained. To take into account this trade-off as a combined effect of block VQ and Kalman filtering, effective block length is determined by evaluating the IS distance [29] defined as follows.

$$\mathrm{IS}(\mathbf{a},\hat{\mathbf{a}}) = \frac{(\mathbf{a}-\hat{\mathbf{a}})^T\mathbf{R}(\mathbf{a}-\hat{\mathbf{a}})}{\mathbf{a}^T\mathbf{R}\mathbf{a}} \tag{34}$$

where $\mathbf{a}$ and $\hat{\mathbf{a}}$ are the direct LPC filter coefficients of the reference and enhanced signals, respectively. R is the reference autocorrelation matrix. IS distance is adopted as it measures the spectral difference independent of energy impact, which is suitable for the proposed energy-independent training and tracking process. Fig.4 shows the IS scores obtained for both speaker-dependent (SD) and speaker-independent (SI) assessments with various block length settings. It is observed in both evaluations that lowest IS score, which indicates smallest tracking error, is achieved with the effective block length around 96ms.

It is worth mentioning that, in practice, the effect of spectral envelope distortion is quite different in various noise conditions. To cover all possible observed features, it is desirable to train a fairly large corpus with a great number of clusters.

In our experiment, an effective solution to reduce the size of training set is to apply a rough estimation of noise spectrum for each frame, and subtract it from the observed spectrum before feature extraction. This step minimizes the effect of noise corruption in various color noise environments as well as various SNR conditions, and hence effectively reduce the cluster size requirement. The simplest approach is to estimate and update the noise spectrum during silent period, using a voice activity detector (VAD). More complicated continuous noise tracking algorithm is also possible.

C. Tracking of Spectral Envelopes

To validate the design of the energy-independent dynamics tracking scheme, comparison of energy-normalized short-time envelope spectra and spectrograms is performed in this subsection. Fig.5 shows the short-term envelope spectra obtained from clean, noisy (white noise, 0dB), pre-cleaned, and Kalman tracked LSFs. It is noticed that unlike conventional methods, no isolated spectral peaks and accurate formants are observed from the spectral envelope which is tracked by Kalman filter. Fig.6 and Fig.7 illustrate the temporal correlation between consecutive spectral envelopes by showing its time-frequency trajectories. In contrast to the spectral fluctuations observed either in noisy or pre-cleaned spectrograms, very smooth and natural envelope evolution is noticed in the tracked trajectory. Meanwhile, common problems in conventional methods such as “musical tones” are completely avoided.

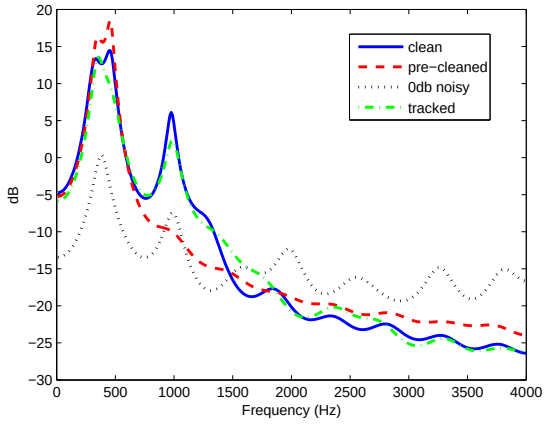

Fig. 5: Short-time envelope spectra of original clean, noisy (white noise, SNR = 0dB), pre-cleaned and Kalman tracked speech segments

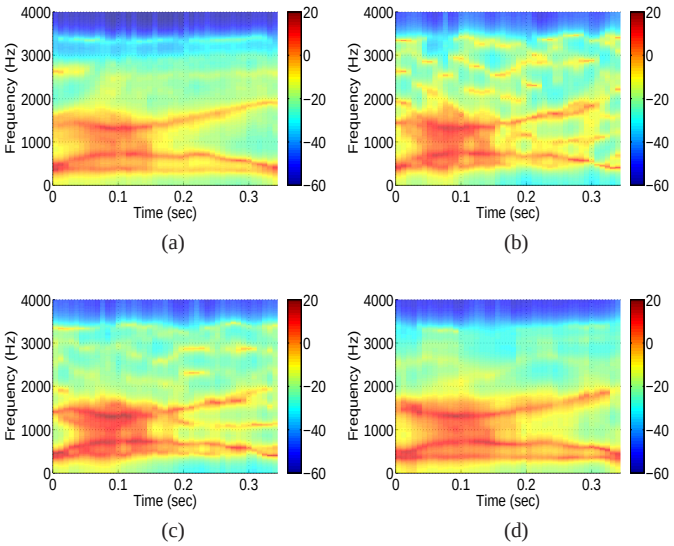

Fig. 6: 2D view of energy-normalized (a) original clean, (b) noisy (white noise, SNR = 0dB), (c) pre-cleaned and (d) Kalman tracked spectral envelope trajectories

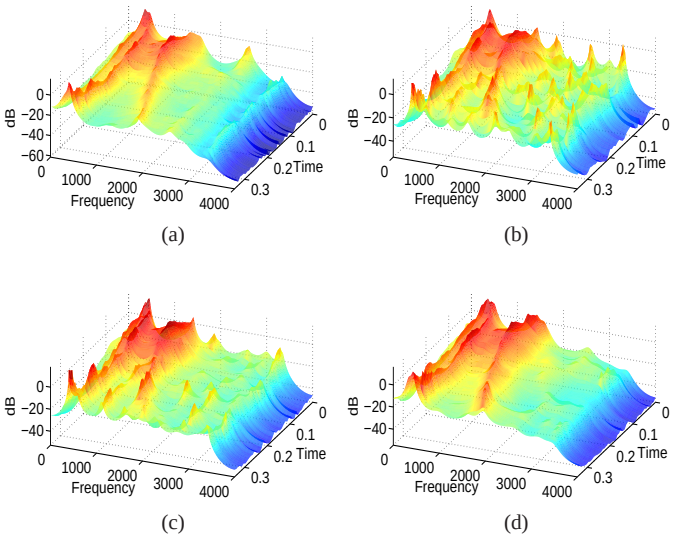

Fig. 7: 3D view of energy-normalized (a) original clean, (b) noisy (white noise, SNR = 0dB), (c) pre-cleaned and (d) Kalman tracked spectral envelope trajectories

D. Computational Complexity

Computational cost for the real-time implementation of the proposed system is discussed in this subsection. There are three major issues that constitute the core computational load in the online enhancement system. First, for each analysis frame, the block codebook searching requires to calculate the distortion (in general, a quadratic measure) between the observation LSF block and all block entries in the codebook. It requires $(P \times P + P)KJ$ multiplications, where P is the order of LSF, K is the block length and J is the total number of entries in the codebook. However, the computational cost can be significantly reduced as the distortion for previous frames in this block has been calculated in previous blocks and hence can be reused with a memory system. Furthermore, the diagonalized sensitivity matrix in (24) shows that the distortion can be calculated in a WMSE form. As a result, in practice, only $2PJ$ multiplications are needed. Second, LSF estimate of current frame is computed online through Kalman filtering with length K. To achieve the iterative update as in section II-B3, a Kalman smoother is desired to obtain accurate estimate of earlier frames in this block based on all

the frames inside this block (including future frames). The online implementation of Kalman smoother is relatively computational demanding, and one way to reduce the computation is that the initial state is derived using (32), in which case only the estimate of last frame is needed and the original Kalman filter is sufficient. Third, the spectrum matching based pitch searching algorithm can also be computational intensive when a full search of possible human pitch range is performed. In practice, similar to the methods in [9] and [11], maximum rate of change of pitch frequency between consecutive frames is defined. Consequently, given the pitch value of last frame, the current search could be limited to a small portion of full range and hence the computation is significantly reduced. In summary, in the proposed system, the setting of the system parameters is relatively flexible, and it can be fine-tuned for applications with different computational requirements.

IV. Performance Evaluation

Performance of the complete speech enhancement system is evaluated in this section. Both SD and SI assessments are conducted. In SD assessment, clean speech is taken from the IEEE sentence database which contains 8KHz-sampling sentences spoken by a single male speaker. In SI assessment, clean speech is taken from TIMIT corpus, which contains a mix of utterances (both male and female speakers, from different dialect regions) with 16KHz-sampling (down-sampled to 8KHz in this experiment). In both experimental settings, noise is taken from the NOISEX-92 database. Three types of stationary noise, namely, Gaussian white noise, car interior noise and F16 cockpit noise and two types non-stationary noise, namely, babble noise and factory noise are employed. Clean speech is manually corrupted by additive noise at SNR level of 0dB, 5dB and 10dB. The parallel training set is a pool of mixed noisy features at different SNRs, along with their true clean representations (many-to-one mapping). The total length of training data is approximately 40 minutes for both SD and SI assessments. Separate testing data (different from training, approximately 5 minutes) for both assessments are also extracted from their own corpus. The proposed HNM-based method with Kalman tracking (denoted as KF_HNM) as well as the original HNM-based approach [10] (denoted as HNM) are compared with four variants of STSA-based methods, including the classical STSA method [1], the LSA method [2], the LSA_SPU) method [7], and the OM_LSA method [3]. The degraded speech without enhancement is denoted as NOISY. To simulate the frequency characteristics of telephony communication, all speech and noise signals are filtered by the modified intermediate reference system (IRS) filters as suggested in ITU-T P.862 [30]. Other system parameters are determined by experimental validations as follows. The block and frame duration are 96ms and 32ms, respectively. Both block shift and frame shift are 8ms. The order of LPC analysis is 18. The total number of clusters for SD and SI training are 64 and 256, respectively.

To evaluate the re-synthesized speech as a combined effect of estimated pitch, spectral gain as well as the tracked spectral envelope, short-time magnitude spectra and spectrograms of the final synthetic speech are studied. Fig.8 and Fig.9 demonstrate the improvement of spectrum amplitude. Fig.8(a) shows the noisy (0dB, white noise) input short-time speech spectrum, with many weak harmonics dominated by noise in high frequency region. Fig.8(b) and Fig.8(c) show that the LSA and OM_LSA methods are able to preserve strong harmonics and at the same time suppress the average noise floor. However, it is observed that spectral shape is further distorted and the harmonic structure is damaged. Fig.8(d) shows the magnitude spectrum constructed by HNM re-synthesis, using pre-cleaned spectral envelope. It is evident that the harmonic structure is restored to a large extent. Nevertheless, the harmonic magnitudes are not well-fitted with the originals due to the mismatched spectral envelope. Fig.8(e) shows the re-synthesized spectrum using tracked spectral envelope. It is observed that the harmonic magnitudes are close to the originals, and even some high-frequency weak harmonics are restored. The spectrograms shown in Fig.9 demonstrate the effectiveness of the proposed method in time-frequency perspective. Fig.9(c) and Fig.9(d) show the trade-off between residual noise suppression and spectral distortion. Fig.9(e) illustrates the restored harmonic structure. Fig.9(f) demonstrates the further enhancement by showing the extended harmonic structure as well as the improved spectral envelope.

A. Objective Evaluation

Three objective evaluation tools, namely IS, LSD and perceptual evaluation of speech quality (PESQ) measures [31] are employed to assess the proposed speech enhancement from different perspectives. Experimental results for both SD and SI tests are shown in Table I and Table II, respectively. As previously discussed, IS measure compares the dissimilarity between two energy-normalized spectral envelopes at their true energy levels. Therefore, it is convenient and effective to adopt this measure to evaluate the improvement in spectral envelope estimation of the proposed design. It is observed from the IS scores that white noise results in largest spectral envelope distortion as it distorts the spectrum uniformly. The LSA method generally causes smallest spectral distortion among conventional approaches. The original HNM method uses the pre-cleaned envelope so that the result is close to corresponding selected pre-clean algorithm, and hence is omitted in Table I and Table II. The KF_HNM method causes least spectral envelope distortion in all SNR settings, which indicates the tracked spectral envelope is closest to the original in various conditions.

Besides, LSD measure compares the difference between log-scale magnitude spectra of noisy and enhanced speech. It is defined as:

$$\mathrm{LSD}(S(\omega),\hat{S}(\omega)) = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{S(\omega)}{\hat{S}(\omega)}\right]^2 d\omega} \quad (35)$$

To emphasize on the overall spectral improvement as a combined effect of spectral envelope plus gain decomposition, $S(\omega)$ and $\hat{S}(\omega)$ represent the gain-normalized envelope spectra rather than the magnitude spectra in this evaluation. Observed

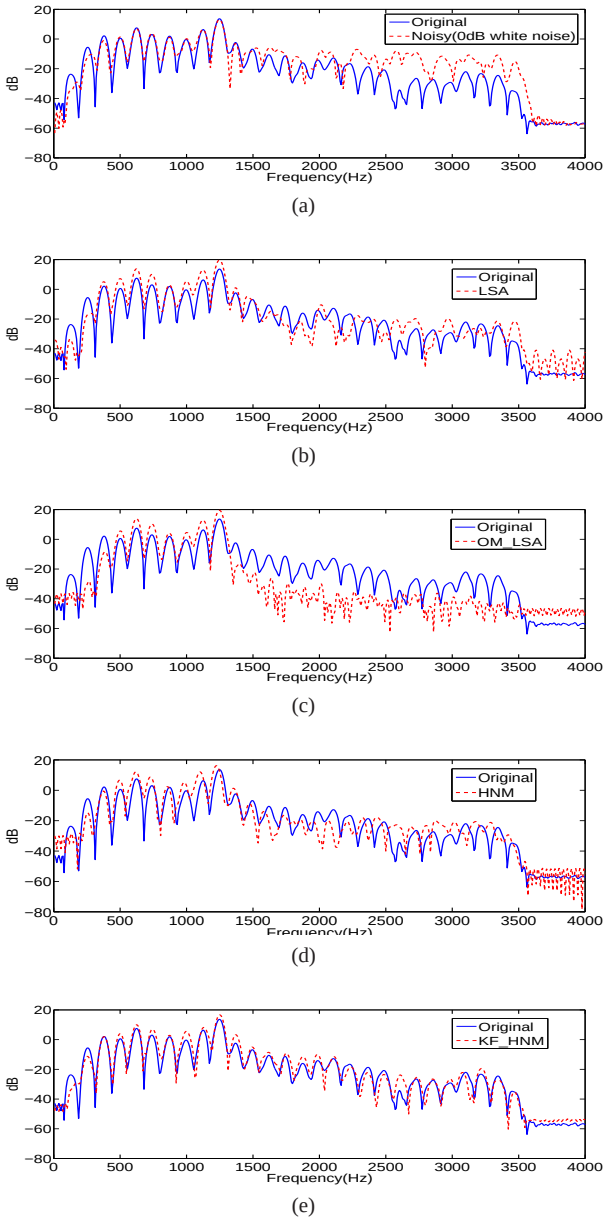

Fig. 8: Comparison of short-time original clean magnitude spectrum versus (a) noisy, (b) LSA, (c) OM_LSA, (d) HNM, and (e) KF_HNM processed magnitude spectra

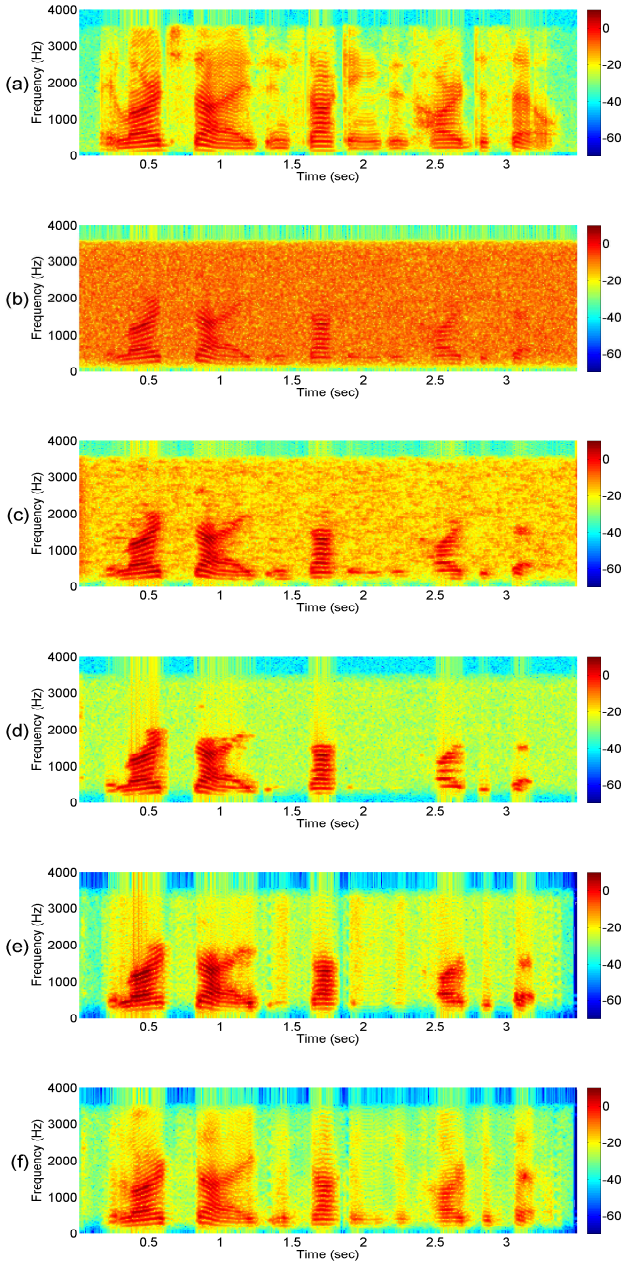

Fig. 9: Spectrograms of (a) original clean, (b) noisy, (c) LSA, (d) OM_LSA, (e) HNM, and (f) KF_HNM processed speech segments

from the LSD results, it is evident that the proposed envelope-plus-gain decomposition is effective as it correlates well with the IS score and an average of over 2dB improvement is achieved using tracked envelope plus pre-cleaned gain in various conditions.

In contrast to previous two measures, standard PESQ measure examines the overall quality of the re-synthesized speech as a combined effect on the complete system design. In addition, it takes into account the psychoacoustic properties. It is noted that an average of around 0.7 point improvement over degraded speech (without enhancement) and an average of around 0.3 point improvement over the best conventional method are achieved in both SD and SI tests in various conditions. To summarize, the proposed method outperforms conventional methods in terms of objective measures in nearly all (different SNR and noise environments) conditions. In particular, the performance gain increases as the SNR decreases.

B. Subjective Evaluation

Subjective evaluation comprises an informal listening test which is designed to follow the procedure suggested in [5]. A total of 10 car-noise-corrupted speech sentences (5 by male and 5 by female) are randomly extracted from the SD testing set at SNR of 5dB. 20 listeners are instructed to successively attend to and rate the enhanced speech (also the noisy speech for benchmarking) on signal distortion

TABLE I: Objective evaluation results of speaker dependent experiment

		Speaker-dependent Experiment								
		IS Distance			LSD (in dB)			PESQ (out of 4.5)		
Noise		Input SNR			Input SNR			Input SNR		
Type	Method	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Gaussian	NOISY	16.39	8.55	4.25	10.51	9.22	7.88	1.68	1.92	2.21
White	STSA	15.15	3.15	1.48	8.77	7.12	6.24	2.08	2.36	2.71
Noise	LSA	6.88	1.95	1.11	8.37	6.69	5.72	2.09	2.41	2.73
	LSA_SPU	14.66	3.26	2.06	8.67	7.38	6.89	1.95	2.22	2.61
	OM_LSA	17.89	6.58	3.71	9.18	8.18	7.68	1.73	2.17	2.60
	HNM	–	–	–	–	–	–	2.21	2.39	2.65
	KF_HNM	**2.42**	**1.53**	**0.84**	**6.54**	**5.42**	**4.64**	**2.42**	**2.61**	**2.82**
Car	NOISY	3.01	2.25	1.72	6.25	5.38	5.56	1.98	2.25	2.55
Interior	STSA	3.35	2.85	2.38	6.78	6.23	5.71	2.24	2.58	**2.92**
Noise	LSA	2.53	1.98	1.74	6.03	5.42	4.93	2.25	2.55	2.86
	LSA_SPU	4.05	3.65	3.02	7.31	6.94	6.42	2.14	2.46	2.90
	OM_LSA	6.06	5.34	3.70	8.03	7.63	6.83	1.88	2.28	2.72
	HNM	–	–	–	–	–	–	2.29	2.61	2.70
	KF_HNM	**2.01**	**1.76**	**1.54**	**5.74**	**5.12**	**4.58**	**2.45**	**2.74**	2.88
F16	NOISY	4.58	2.84	1.93	8.32	7.28	5.59	1.81	2.09	2.39
Cockpit	STSA	5.49	2.32	1.39	7.84	6.53	5.45	2.19	2.53	2.86
Noise	LSA	2.50	1.37	1.21	7.09	5.89	5.56	2.23	2.56	2.87
	LSA_SPU	6.17	2.59	1.81	8.03	6.92	6.62	2.02	2.42	2.82
	OM_LSA	9.26	4.06	2.58	8.69	7.65	7.50	1.69	2.28	2.65
	HNM	–	–	–	–	–	–	2.31	2.59	2.72
	KF_HNM	**1.97**	**1.20**	**0.96**	**6.41**	**5.47**	**4.59**	**2.47**	**2.71**	**2.88**
Babble	NOISY	4.45	2.33	1.93	10.00	8.89	7.47	1.92	2.23	2.44
Noise	STSA	6.24	4.57	4.12	10.93	8.94	7.87	2.06	2.43	2.65
	LSA	5.47	3.09	2.89	10.76	8.59	7.75	2.02	2.40	2.61
	LSA_SPU	10.17	6.17	6.05	11.91	9.58	8.52	1.94	2.23	2.52
	OM_LSA	11.39	8.78	8.12	11.79	9.89	8.81	1.83	2.09	2.43
	HNM	–	–	–	–	–	–	2.11	2.45	2.55
	KF_HNM	**3.22**	**2.21**	**1.90**	**7.19**	**5.35**	**4.79**	**2.33**	**2.53**	**2.68**
Factory	NOISY	5.21	3.37	2.09	9.74	8.78	6.36	1.74	2.07	2.38
Noise	STSA	7.49	5.43	3.92	9.51	8.57	7.21	2.10	2.43	2.83
	LSA	5.20	2.89	2.13	9.01	8.02	6.30	2.06	2.42	2.84
	LSA_SPU	9.57	8.28	4.73	9.94	8.92	7.42	1.90	2.29	2.81
	OM_LSA	10.26	8.96	5.91	10.07	9.21	8.09	1.59	2.21	2.66
	HNM	–	–	–	–	–	–	2.17	2.40	2.75
	KF_HNM	**2.28**	**2.07**	**0.67**	**5.35**	**4.77**	**3.82**	**2.40**	**2.56**	**2.86**

(SIG)−[1=very unnatural, 5=very natural], background intrusiveness (BAK)−[1=very conspicuous, very intrusiveness, 5=not noticeable], and overall effect using the scale of mean opinion score (OVRL)−[1=bad, 5=excellent]. Fig.10 shows the mean scores of the listening test for the three scales.

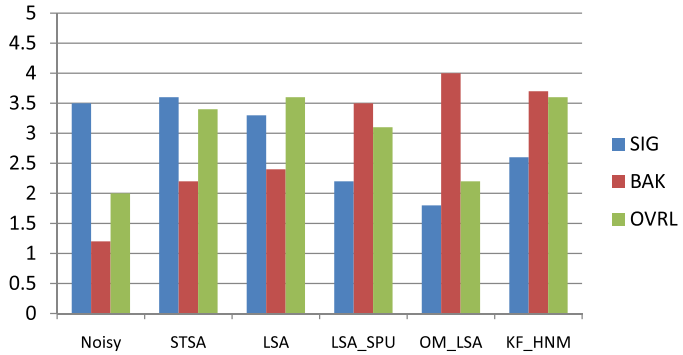

Fig. 10: Subjective evaluation results of speaker dependent experiment

Mean scores in Fig.10 demonstrate the trade-off between signal distortion and background intrusiveness for various enhanced signals. Higher SIG scores indicate that the STSA and LSA methods preserve relatively higher level of signal quality. However, lower BAK scores reflect the downside, which is the negative impact of annoying artifacts. Conversely, superior BAK scores show that the LSA_SPU and OM_LSA methods did a good job in noise suppression. Nevertheless, poor SIG scores reveal the severe degradation of signal quality. The OVRL scores suggest that more participants prefer the former approaches. The proposed KF_HNM method serves as a compromise between the above-mentioned trade-off. On one hand, it is observed that the SIG score of the KF_HNM method is lower than the STSA and LSA methods. Since the enhanced speech is constructed by means of re-synthesis, signal degradation mainly comes from the modeling and estimation error of the proposed design. On the other hand, the BAK score of the KF_HNM method is much higher than those of the STSA and LSA methods, owing to its automatic noise removal capability. It is noted that the BAK score is slightly lower than that of the OM_LSA method. The potential reason is that several re-synthesis artifacts are perceived as noise by the subject and hence result in reduced BAK score. The OVRL score of the KF_HNM method is among the best for all evaluating methods. Although the mean OVRL scores of the LSA and KF_HNM methods are close, individual assessments and comments vary for each subject. Based on the feedbacks from participants, some of them prefer the KF_HNM method as it successfully removes the annoying tonal effect, whereas

TABLE II: Objective evaluation results of speaker independent experiment

		Speaker-independent Experiment								
		IS Distance			LSD (in dB)			PESQ (out of 4.5)		
Noise		Input SNR			Input SNR			Input SNR		
Type	Method	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Gaussian	NOISY	8.05	5.82	3.81	8.95	8.01	7.06	1.51	1.79	2.13
White	STSA	8.72	5.32	3.16	8.51	7.48	6.82	1.90	2.34	2.70
Noise	LSA	4.67	2.99	1.89	7.92	6.91	6.15	1.92	2.34	2.71
	LSA_SPU	8.60	5.37	3.51	8.44	7.68	7.38	1.79	2.15	2.56
	OM_LSA	9.68	6.44	5.89	8.71	8.43	8.35	1.48	1.87	2.21
	HNM	–	–	–	–	–	–	2.09	2.39	2.65
	KF_HNM	**3.42**	**2.53**	**1.64**	**6.04**	**5.42**	**4.84**	**2.21**	**2.53**	**2.72**
Car	NOISY	3.18	2.55	2.04	7.40	6.50	5.55	1.73	2.06	2.40
Interior	STSA	3.16	2.83	2.56	6.97	6.65	6.37	2.25	2.56	**2.89**
Noise	LSA	2.72	2.31	2.11	6.82	6.20	5.71	2.19	2.52	2.83
	LSA_SPU	3.51	3.63	4.02	7.40	7.39	7.18	1.96	2.38	2.76
	OM_LSA	5.60	5.29	5.10	8.34	8.12	8.04	1.53	2.01	2.42
	HNM	–	–	–	–	–	–	2.29	2.53	2.65
	KF_HNM	**1.95**	**1.75**	**1.64**	**5.32**	**5.07**	**4.88**	**2.40**	**2.64**	2.80
F16	NOISY	2.78	2.21	1.49	7.78	7.11	6.17	1.67	1.97	2.32
Cockpit	STSA	4.40	2.95	2.23	8.00	7.21	6.63	2.13	2.47	2.79
Noise	LSA	2.50	1.71	1.32	7.33	6.55	5.91	2.12	2.47	**2.80**
	LSA_SPU	4.53	3.28	3.10	8.14	7.56	7.28	1.89	2.25	2.71
	OM_LSA	6.15	5.76	5.64	8.82	8.51	8.04	1.54	1.93	2.33
	HNM	–	–	–	–	–	–	2.28	2.39	2.65
	KF_HNM	**1.91**	**1.57**	**1.30**	**5.71**	**5.17**	**4.99**	**2.38**	**2.62**	2.78
Babble	NOISY	4.15	2.17	1.74	8.71	7.48	6.25	1.59	2.04	2.36
Noise	STSA	6.01	3.47	3.12	9.10	8.35	7.01	1.76	2.27	2.54
	LSA	5.74	2.98	2.51	8.93	7.84	6.75	1.68	2.24	2.50
	LSA_SPU	6.47	3.69	3.55	9.81	8.24	7.34	1.55	2.16	2.39
	OM_LSA	7.19	5.21	4.97	10.01	8.79	7.68	1.50	2.16	2.40
	HNM	–	–	–	–	–	–	1.87	2.25	2.50
	KF_HNM	**3.42**	**2.15**	**1.71**	**7.04**	**6.15**	**5.01**	**2.13**	**2.48**	**2.59**
Factory	NOISY	3.75	2.84	1.92	8.08	7.58	6.25	1.55	1.86	2.32
Noise	STSA	4.25	3.78	2.96	8.34	8.06	6.95	1.92	2.18	2.71
	LSA	3.20	2.81	2.27	7.81	7.67	6.74	1.95	2.21	**2.75**
	LSA_SPU	5.71	4.28	3.45	8.54	8.29	7.31	1.71	2.07	2.67
	OM_LSA	6.29	4.96	3.65	8.77	8.61	7.41	1.61	1.84	2.51
	HNM	–	–	–	–	–	–	2.01	2.21	2.67
	KF_HNM	**2.52**	**2.11**	**1.27**	**7.12**	**6.34**	**5.98**	**2.10**	**2.30**	2.73

some prefer the LSA method as they are more sensitive to the noise-like component caused in synthetic speech (sounds a bit hoarse).

C. Discussion

To summarize the findings in both objective and subjective evaluations, it is observed that the proposed KF_HNM method achieves obvious improvement in various objective assessments. While subjective listening test results show relatively diverse opinions, which is not necessarily correlated with the objective measures. Listening test results suggest that the proposed method is preferred by some subjects, yet still several shortcomings exist. The major reason is that several perceivable distortion is occasionally observed in the synthetic speech. This is mainly due to the error in modeling, clustering, and estimation strategies in adverse conditions. In summary, the proposed method offers a new direction for speech enhancement and it already exhibits many advantages (such as musical tone removal and harmonic structure restoration). In addition, owing to its flexible decomposition, this approach can be improved in many perspectives. In future research work, we suggest investigating three issues, namely 1) forming more robust and informative feature representations for various noisy observations, 2) improving the corresponding noise-robust clustering and identification strategies, and 3) incorporating more sophisticated dynamic speech modeling techniques to accurately model the target speech in transient periods.

V. Conclusion

We have proposed a new speech enhancement system which explores the inherent time-frequency characteristics of speech. HNM based analysis-synthesis framework is employed to take advantage of the harmonic structure of speech while a dynamics tracking scheme is incorporated to exploit the temporal correlation between spectral envelope trajectories. The proposed system provides robust parameter estimation algorithms, and hence operates consistently good in various noisy conditions. Furthermore, it offers flexibility in independent parameter adjustment and hence could be optimized according to various noisy conditions. Both objective and subjective evaluation results demonstrate the effectiveness of the proposed system.

Acknowledgment

The authors wish to thank the anonymous reviewers for their careful reading and valuable comments that improved the quality of this paper.

REFERENCES

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

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoust., Speech and Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[4] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," *Proceedings of the International Conference on the Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, p. I, May 2006.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: Taylor & Francis Group, 2007.

[6] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[7] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Signal Processing*, vol. 75, no. 2, pp. 151–159, 1999.

[8] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 157–160, Mar. 2005.

[9] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1194–1203, May 2007.

[10] R. F. Chen, C. F. Chan, H. C. So, J. Lee, and C. Y. Leung, "Speech enhancement in car noise environment based on an analysis-synthesis approach using harmonic noise model," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4413–4416, Apr. 2009.

[11] R. F. Chen, C. F. Chan, and H. C. So, "Noise suppression based on an analysis-synthesis approach," *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 1539–1543, Aug. 2010.

[12] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[13] T. V. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 383–389, Sep. 1996.

[14] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[15] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, pp. 177–180, Apr. 1987.

[16] J. D. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[17] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[18] E. Zavarehei, S. Vaseghi, and Q. Yan, "Inter-frame modeling of DFT trajectories of speech and noise for speech enhancement using kalman filters," *Speech Communication*, vol. 48, no. 11, pp. 1545–1555, 2006.

[19] T. Eriksson, J. Linden, and J. Skoglund, "Interframe LSF quantization for noisy channels," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 495–509, Sep. 1999.

[20] L. Girin, "Long-term quantization of speech LSF parameters," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 845–848, 2007.

[21] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing,*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[22] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[23] C. F. Chan and E. W. M. Yu, "Improving pitch estimation for efficient multiband excitation coding of speech," *Electronics Letters*, vol. 32, no. 10, pp. 870–872, May 1996.

[24] E. W. M. Yu and C.-F. Chan, "Harmonic+noise coding using improved V/UV mixing and efficient spectral quantization," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 477–480, Mar. 1999.

[25] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[26] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[27] W. Gardner and B. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 367–381, Sep. 1995.

[28] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Transactions on Speech and Audio Process.*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[29] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, Apr. 1993.

[30] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements." *IEEE Transactions on Audio and Electroacoustics*, vol. AU-17(3), no. 10, pp. 225–246, May 1969.

[31] ITU P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation P. 862*, 2000.