Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 22 (2008) 69–83

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement

Qin Yan [a,*], Saeed Vaseghi [a], Esfandiar Zavarehei [a], Ben Milner [b], Jonathan Darch [b], Paul White [c], Ioannis Andrianakis [c]

[a] *School of Computer and Information Engineering, Hohai University, Nanjing 210000, China*
[b] *School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK*
[c] *Institute of Sound and Vibration Research, University Road, Highfield, Southampton S017 1BJ, UK*

Received 13 June 2006; received in revised form 22 May 2007; accepted 15 June 2007
Available online 10 July 2007

Abstract

This paper presents a speech enhancement method based on the tracking and denoising of the formants of a linear prediction (LP) model of the spectral envelope of speech and the parameters of a harmonic noise model (HNM) of its excitation. The main advantages of tracking and denoising the prominent energy contours of speech are the efficient use of the spectral and temporal structures of successive speech frames and a mitigation of processing artefact known as the 'musical noise' or 'musical tones'.

The formant-tracking linear prediction (FTLP) model estimation consists of three stages: (a) speech pre-cleaning based on a spectral amplitude estimation, (b) formant-tracking across successive speech frames using the Viterbi method, and (c) Kalman filtering of the formant trajectories across successive speech frames.

The HNM parameters for the excitation signal comprise; voiced/unvoiced decision, the fundamental frequency, the harmonics' amplitudes and the variance of the noise component of excitation. A frequency-domain pitch extraction method is proposed that searches for the peak signal to noise ratios (SNRs) at the harmonics. For each speech frame several pitch candidates are calculated. An estimate of the pitch trajectory across successive frames is obtained using a Viterbi decoder. The trajectories of the noisy excitation harmonics across successive speech frames are modeled and denoised using Kalman filters.

The proposed method is used to deconstruct noisy speech, de-noise its model parameters and then reconstitute speech from its cleaned parts. Experimental evaluations show the performance gains of the formant tracking, pitch extraction and noise reduction stages.

Keywords: HNM; Kalman; Formant

[*] Corresponding author.
E-mail addresses: yanqin@ieee.org (Q. Yan), Saeed.Vaseghi@brunel.ac.uk (S. Vaseghi), Esfandiar.Zavarehei@brunel.ac.uk (E. Zavarehei), bpm@cmp.uea.ac.uk (B. Milner), jd@cmp.uea.ac.uk (J. Darch), prw@isvr.soton.ac.uk (P. White), ia@isvr.soton.ac.uk (I. Andrianakis).

0885-2308/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2007.06.002

1. Introduction

Enhancement of noisy speech improves the quality and intelligibility of voice communication in noisy environments such as for mobile and hands-free phones used in noisy public venues as in busy streets, moving cars and trains, noisy conference halls, cafés, noisy shops/markets, airports, factory floors, etc. Since the method proposed in this paper is based on the explicit estimation of the parameters of a linear prediction (LP) model and a harmonic noise model (HNM) of speech, it also has applications in speaker identification/verification and speech recognition in noise. However, the focus of this paper is on speech enhancement.

A desirable property of speech enhancement is that it should not replace the noise by processing artefacts which may be just as detrimental to the quality or intelligibility of speech as the original noise. A main advantage of the method proposed here is a mitigation of a commonly occurring processing artefact, known as 'musical noise' or 'musical tones', composed of random short duration bursts of narrowband (tonal) noise that are often an undesirable byproduct of noise reduction methods.

In the development of the speech enhancement method described here, the main objective was to utilize the statistical models of the trajectories of the prominent energy contours of speech, namely those of formants and harmonics (Lim and Oppenheim, 1978; Stylianou, 1996). This is motivated by the observations that formants and HNM parameters, which represent the peak energy contours of speech, usually have relatively higher than average signal to noise ratio (SNR) and that speech can be reconstructed from these parameters. It may be that the robustness of human speech recognition is partly due to the structure of speech where valuable information resides in the high SNR parts of speech and in particular in the formants and the harmonics of excitation.

The proposed method integrates a formant-tracking linear prediction (FTLP) model of spectral envelope with a HNM of excitation. LP and HNM (Lim and Oppenheim, 1978; Stylianou, 1996; Vaseghi, 2006) are the two main methods for modeling speech waveforms; they offer complementary advantages; LP model provides a good fit for the spectral envelope of speech whereas HNM is good at modeling the details of the harmonic plus noise structure of speech excitation.

For noisy speech enhancement the proposed approach is different to conventional methods such as spectral subtraction (Vaseghi, 2006; Boll, 1979) minimum mean squared error spectral amplitude (Ephraim and Van Trees, 1995; Ephraim, 1985); the variants of Wiener filters (Hansen and Clements, 1987; Sameti et al., 1998; Ephraim, 1992; Ephraim et al., 1989; Chen et al., 2000) and speech enhancement methods based on linear prediction models (Lim and Oppenheim, 1979). In conventional speech enhancement methods often individual spectral samples are modeled in isolation without fully utilizing the information on the wider spectral-temporal structures that may be used to good effect in the de-noising process to obtain improved speech enhancement results.

Speech processing systems normally segment speech into a sequence of frames with a duration of about 20–30 ms. Two major issues in speech signal processing are: (1) the modeling of the intra-frame correlation of time (or frequency) samples within each speech frame and (2) the modeling of the inter-frame correlation of speech samples across successive frames of speech. The proposed FTLP-HNM model, with Viterbi trackers and Kalman filters, provides a suitable framework for modeling the intra-frame and inter-frame non-stationary temporal variations of speech parameters across successive speech frames and this can reduce the errors and uncertainty in estimation of speech model parameters.

The FTLP model obtains enhanced estimates of the LP parameters of speech along the formant trajectories. Formants are the resonances of the vocal tract and their trajectories describe the contours of energy concentrations in time and frequency. Although formants are mainly defined for voiced speech, characteristic energy contours also exist for unvoiced speech as concentrations of energy at relatively higher frequencies. In the context of noisy speech processing, the use of speech features at formants is particularly interesting because at the formants the SNRs are relatively high and furthermore much of the discriminative information regarding phonemic labels and some of the speaker characteristics are encoded in the spectral features at formants.

In this paper, HNMs are used to model the trajectories of the excitation of LP model. This makes good sense given that LP-based speech coders (such as mobile phone speech coders) also use a combination of periodic and non-periodic signals for speech excitation. HNM is an established method particularly in speech and music coding and text to speech synthesis (Stylianou, 1996). The main issues in HNM are voiced/unvoiced classification and the estimation of fundamental frequency (pitch) value, the harmonic amplitudes and a model of the noise component of speech excitation.

For pitch estimation in noise a modified version of Griffin's method (Griffin and Lim, 1988) is introduced. The method is based on frequency domain estimation of the pitch from discrete Fourier transform (DFT) of speech and allows the use of estimates of SNRs at the harmonics for improved performance. The criterion used for pitch estimation is based on searching for the peaks of the spectral energy at the harmonics of speech. The pitch estimate is derived through searching a grid in a range of proposed pitch values. Errors in pitch estimation are corrected by the use of a Viterbi estimation algorithm. Further smoothing and improvement in pitch trajectory estimates may be obtained by the use of a Kalman filter.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the proposed FTLP-HNM model estimation method. Section 3 presents a review of the method for extraction of formant features and introduces the probability model used for formant features. Section 4 introduces HNM of excitation and presents a new method for pitch estimation over harmonics. Section 5 discusses Kalman filters and their application in the modeling and smoothing of the trajectories of formants and the denoising harmonics of speech excitation. Section 6 describes performance measures for speech enhancement and presents evaluation results. Finally, Section 7 concludes the paper.

2. An overview of formant-tracking LP model with HNM of excitation

The proposed FTLP-HNM for enhancement and de-noising of noisy speech is illustrated in Fig. 1 and consists of the following sections:

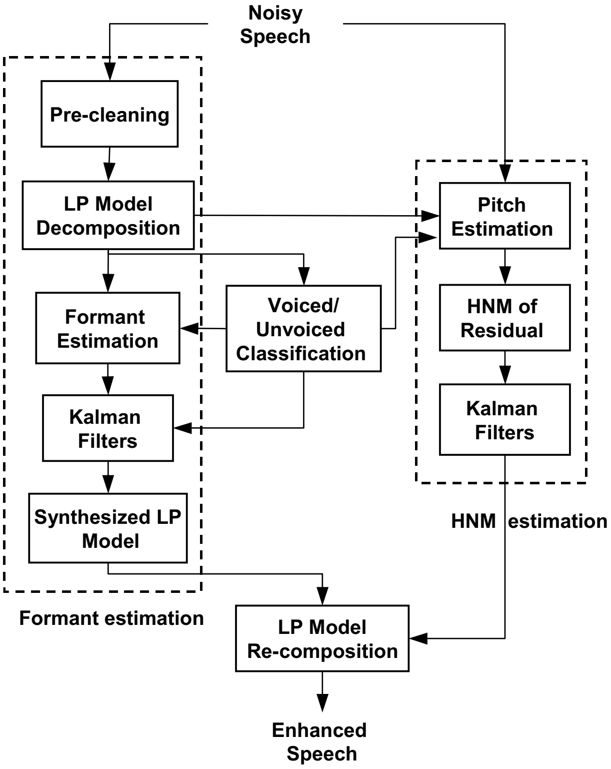

Fig. 1. Overview of the FTLP-HNM model for enhancement of noisy speech.

(1) A pre-cleaning module for de-noising speech prior to estimation of the LP model and formant parameters.
(2) A formant-tracking method incorporating Viterbi decoders and Kalman filters for tracking and smoothing the temporal trajectories of significant formants and poles of the LP model of speech.
(3) A pitch extraction method incorporating Viterbi decoders and Kalman filters for pitch tracking and smoothing.
(4) A harmonic noise model estimation method using Kalman filters for modeling and denoising the temporal trajectory of noisy excitation.

The z-transform of a linear prediction model of speech $X(z,m)$ may be expressed as

$$X(z,m) = E(z,m)V(z,m) \tag{1}$$

where $E(z,m)$ is the z-transform of the excitation and $V(z,m)$ is the z-transform of a LP model of the combined effect of the vocal tract, the glottal pulse and the lip radiation. The vocal tract model $V(z,m)$ can be expressed as a cascade combination of a set of second order resonators and a first order model as

$$V(z,m) = G(m)\frac{1}{1+r_0(m)z^{-1}}\prod_{k=1}^{P/2}\frac{1}{1-2r_k(m)\cos(\varphi_k(m))z^{-1}+r_k^2(m)z^{-2}} \tag{2}$$

where $r_k(m)$ and $\phi_k(m)$ are the time-varying radii and the angular frequencies of the poles of the LP model respectively, $P+1$ is the LP model order and $G(m)$ is the gain of the LP model. In Eq. (2) speech is modeled by a cascade of time-varying second order resonator models of the formants and a first order model of the slope of speech spectrum. For voiced speech, the resonators are associated with the formants of speech. For unvoiced speech, the second order resonators model the energy concentrations of speech.

The speech excitation can be modeled as a combination of the harmonic and the noise as

$$e(m) = \sum_{k=1}^{L(m)} A_k(m)\cos(2\pi(kF_0(m)+\varDelta_k(m))m+\varphi_k(m))+v(m) \tag{3}$$

where $F_0(m)$ is the time-varying fundamental frequency of speech excitation, $A_k(m)$ are the magnitude of excitation harmonics, $\varphi_k(m)$ are the phase of harmonics, $\varDelta_k(m)$ is the deviation of the kth harmonic from the nominal value of kF_0 and $v(m)$ is the noise part of the excitation. It is assumed that phase distortion is inaudible. In the following sections, we explain the estimation of FTLP model coefficients and the HNM parameters of the excitation.

3. Estimation of a formant-tracking LP model from noisy speech

The spectral envelop of speech is modeled by the frequency response of the LP model. This section describes the formant-tracking LP model (FTLP) estimation process composed of three stages of; pre-cleaning of speech spectrum, formant classification and Kalman filters for smoothing formant trajectories.

3.1. Initial-cleaning of spectral amplitudes of noisy speech

Before formant estimation, pre-cleaning of noisy speech, is accomplished through estimation of the spectral amplitude of speech using a MMSE spectral amplitude estimation method (Ephraim and Van Trees, 1995). The MMSE method of estimation of the spectral amplitude of speech is a Bayesian estimation method employing a mean squared error cost function. In the MMSE formulation of Ephrahim and Mallah (Ephraim, 1985) used here, it is assumed that the prior probability density function (pdf) of the complex spectrum of clean speech is Gaussian. This assumption leads to a Rayleigh pdf for the magnitude spectrum of clean speech and a uniform pdf for its phase. It is further assumed that the complex spectrum of noise has a Gaussian pdf (Ephraim, 1985).

After pre-cleaning the spectral amplitude of speech is converted to a correlation function from which an initial estimate of the LP model of speech is obtained. The poles of the LP model are obtained through a factorization of the LP model polynomial using a rooting function. A formant tracker is then used to obtain an improved estimate of the LP model parameters as described in the next section.

3.2. HMM-based formant tracking

The poles, obtained from factorizing the LP model polynomial of pre-cleaned speech, are the formant candidates represented as formant feature vectors, $\boldsymbol{V}_k$ comprising the frequency, F_k, bandwidth, B_k and magnitude, M_k, of the resonance at formants together with their velocity derivatives as

$$\boldsymbol{V}_k = [F_k, B_k, M_k, \Delta F_k, \Delta B_k, \Delta M_k] \quad k = 1, \ldots, N \tag{4}$$

where the number of formants is typically set to $N = 5$. Velocity derivatives are denoted by Δ and are computed as the slope of the features over time (Rentzos et al., 2003; Weber et al., 2001; Kim and Sung, 2001).

There are two main issues in the accurate modeling of formants; (i) modeling the probability distributions of formants using a probability model such as an HMM or GMM and (ii) tracking and smoothing the trajectory of each formant using a combination of Viterbi classifier followed by Kalman filter for smoothing of the formant trajectories.

The pdfs of the trajectories of the formants can be modeled by HMMs as described in detail in a number of papers to which the interested reader is referred to Rentzos et al. (2003), Weber et al. (2001) and Kim and Sung (2001). Formant HMMs are trained on formant feature vectors of speech obtained from Eq. (4). Given a set of observations of the resonance (pole) frequencies of a speech frame, $\boldsymbol{O}_n$, the maximum likelihood (ML) decoder of the associated formant labels is obtained as

$$\left[\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_N\right] = \arg\max_{F_1, F_2, \ldots, F_N} P(\boldsymbol{O}_n, [F_1, F_2, \ldots, F_N] | \Lambda_m) \quad k = 1, \ldots, N \tag{5}$$

where $\boldsymbol{O}_n$ is obtained from the poles of an LP model of a speech frame and sorted in terms of the increasing frequency, $\widehat{F}_k$ is the ML estimate of the kth formant, Λ_m is an HMM of the formants of phoneme m and $N = 4$–6 is the number of formants. Eq. (6) is implemented using a Viterbi algorithm on HMMs of formants (Rentzos et al., 2003; Weber et al., 2001; Kim and Sung, 2001).

The HMM-based formant classifier may associate two or more formant candidates (poles of LP model) $F_{i(t)}$, with the same formant label k. In these cases, formant estimation is achieved through minimization of a weighted MMSE objective function as (Turunen and Vlaj, 2001)

$$\widehat{F}_k(t) = \arg\min_{F_k(t)} \sum_{i=1}^{N_k(t)} w_{ki}(t) \left(\frac{(F_{ki}(t) - F_k(t))^2}{B_{ki}(t)^2} \right) \quad k = 1, \ldots, N \tag{6}$$

where $F_{ki}(t)$ is the frequency of the ith pole classified as the kth formant, $w_{ki}(t) = P(F_{ki}(t)|\lambda_k)$ is the probability that the ith pole frequency is labeled as the kth formant, λ_k is a Gaussian mixture model of the kth formant and $N_k(t)$ is the total number of poles of the tth speech frame classified as formant k. In Eq. (6) the distance of each pole frequency candidate from the formant estimate is weighted by a probabilistic weight $w_{ki}(t)$ and a perceptual weight $1/B_i^2$ where B_i is the formant bandwidth. Note that $N_k(t)$ is usually one or two. For the case when $N_k(t)$=1 then $\widehat{F}_k(t) = F_{k1}(t)$. For the case when $N_k(t) = 2$, taking the derivative of Eq. (6) with respect to $F_k(t)$ yields a MMSE interpolated estimate of the kth formant at time t as

$$\widehat{F}_k(t) = \frac{\alpha_{k1}}{\alpha_{k1} + \alpha_{k2}} F_{k1}(t) + \frac{\alpha_{k2}}{\alpha_{k1} + \alpha_{k2}} F_{k2}(t) \quad k = 1, \ldots, N \tag{7}$$

where $\alpha_{ki} = w_{ki}(t)/B_{ki}^2(t)$.

3.3. Formant tracking using viterbi decoder with MSE criterion

For speech enhancement applications, in systems where probability models of formants (such as HMM of formants) are not available, the classification of the poles into formant tracks may be achieved with a Viterbi decoder using a minimum squared error (MSE) distance. In such cases a simple yet effective method to take into account the past history of the trajecotry of the process is to obtain the MMSE distance of the current sample from the M past best estimates where $M \geqslant 2$.

In cases where the Viterbi decoding associates two poles with the same formant then a weighted mean of the poles is obtained using Eq. (7) where the weight for each pole may be obtained from $\alpha_{ki} = w_{ki}(t)/B^2_{ki}(t)$ with $w_{ki}(t) = \exp(-\gamma_i(O_i - F_i)^2)$, where γ_i is a control variable, γ_i can be set to the inverse of an estimate of the variance of F_i to produce a measure that is similar to Gaussian probability. The Viterbi-based method of formant tracking, together with HMM of formants, is implemented and evaluated in this paper for speech enhancement.

3.4. Investigation of the effect of noise on formant estimation

The database used to investigate the effect of noise on formants is the Wall Street Journal (WSJ) speech database. Speech is degraded by either car noise (the example used here is a BMW 3 series at 112 kmph) or train noise, with an average SNR in the range from 0 to 20 dB. Formant tracks of clean and denoised speech are obtained via LP-based formant extraction and HMMs reviewed in Section 3.2. To quantify the effects of the noise on formants, a local formant signal to noise ratio measure (FSNR) is introduced (Yan et al., 2004). It is defined as

$$\mathrm{FSNR}(k) = 10\log\left[\sum_{l\in(F_k\pm B_k/2)} X_l^2 \middle/ \sum_{l\in(F_k\pm B_k/2)} N_l^2\right] \tag{8}$$

where X_l is the magnitude spectrum of clean speech, N_l is the magnitude spectrum of noise and F_k and B_k are the frequency and bandwidth of the kth formant. Fig. 2 displays the FSNRs of noisy speech in moving car and train environments. It is evident that the FSNRs are higher than the overall average SNR, which may be a contributing factor to the fact that humans can recognize speech under severe noisy conditions.

The effects of noise on the observations of the formant frequencies of the vowels at different SNRs are shown Fig. 3 which displays the average percentage formant track errors as a function of average SNR. Note that the first formant is most affected by noise due to a greater concentration of the energy of car/train noise at its vicinity.

Fig. 4 illustrates an example of formant tracks of speech in train and car noise at a SNR of 0 dB. The formant tracks are superimposed on LP model spectrogram of clean speech. The formant tracks are obtained from 2D-HMMs. As expected, due to the relatively broader spread of the energy of the train noise in frequency domain compared to that of the car noise, the estimates of formant tracks of noisy speech in the car noise are more robust to noise than those from the train noise. Note that, the first formant track, which is the closest to the concentration of noise in frequency domain, is most affected by noise and this effect on the first formant is even more pronounced in train noise than in car noise. Furthermore, in both cases the first formant F1 is more affected by train/car noise than other formants.

To quantify the affects of noise on speech, an average formant track error measure, defined as the normalized difference between the formant tracks obtained from clean (reference) speech and noisy speech is calculated as follows:

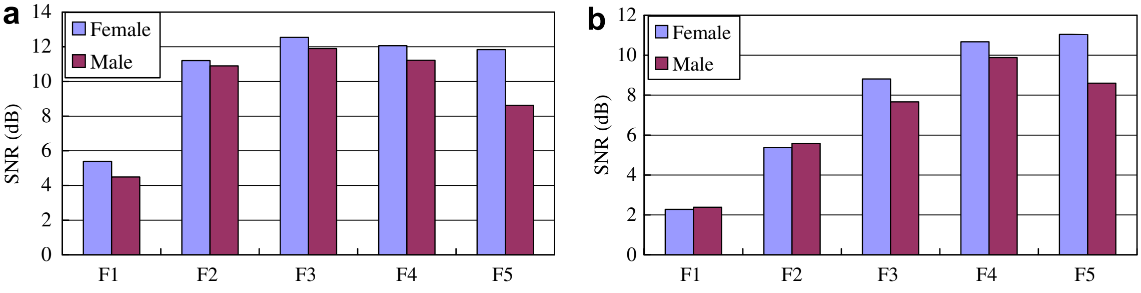

Fig. 2. Variation of speech SNR at formants in (a) car noise (b) train noise at average SNR = 0 dB.

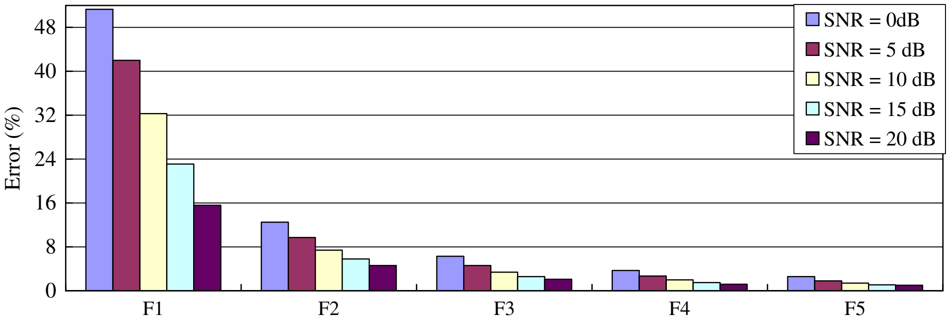

Fig. 3. Average (%) estimation error of speech formant tracks in train noise at different SNRs.

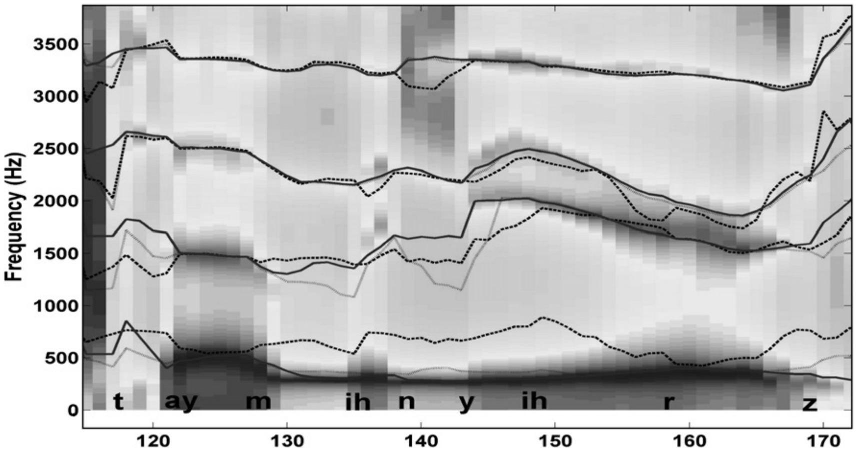

Fig. 4. Illustration of the impact of car/train noise on estimates of the formant tracks of a speech segment "time in years" (SNR = 0 dB). The formant tracks are superimposed on spectrogram of LP model of clean speech. Solid lines: clean speech; dashed lines: formant track in train noise; doted lines: formant track in car noise.

$$E_k = \frac{1}{L}\sum_{m=1}^{L}\left[\frac{\left|F_k(m) - \widehat{F}_k(m)\right|}{F_k(m)}\right] \times 100\% \quad k = 1,\ldots,N \tag{9}$$

where $F_k(m)$ and $\widehat{F}_k(m)$ are the formant tracks of clean and noisy speech respectively, m is frame index and L is the number of frames over which the error is measured. In Fig. 5 the percentage formant track errors are averaged over 135 speech sentences contaminated with train noise. It is evident from Fig. 6 that formant tracking performance degrades with the decreasing SNR.

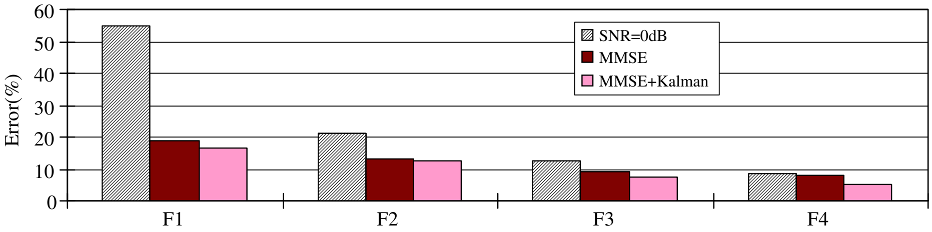

Fig. 5. Average % error of formant tracks in train noise and cleaned speech using MMSE and Kalman filters, the results were averaged over five males.

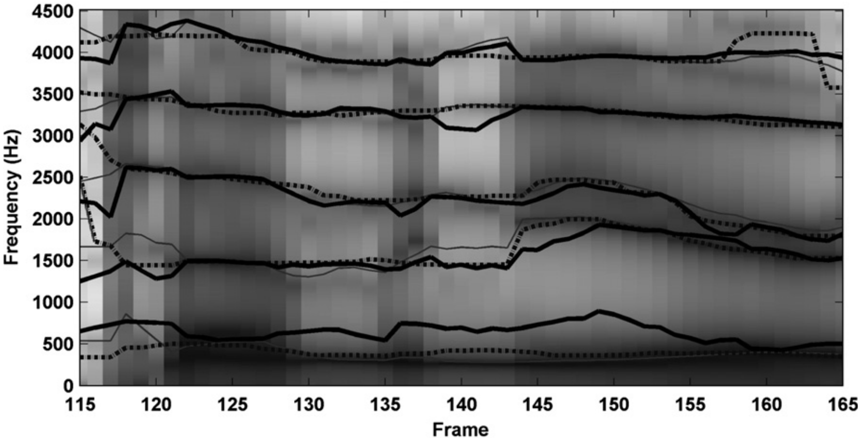

Fig. 6. Comparison of clean formant tracks (thin solid) and cleaned formant tracks (dash dot) and noisy formant tracks (thick solid) for a segment of speech 'time in years'. The background is spectrogram of LP model of clean speech.

3.5. Formant track smoothing with state-dependent Kalman filters

Kalman filters are Bayesian Gaussian–Markov models that (Kalman, 1960) can be used for the modeling and smoothing of a correlated trajectory such as those of the formant trajectories. The application of Kalman filter in formant estimation requires a model of the transition matrix of the formant trajectories and knowledge of the covariance matrices of the process noise and the observation noise.

In this application, the input signals to the Kalman filters are the formant features obtained from classification of the poles of the LP model of pre-cleaned speech as illustrated in the block diagram of Fig. 1. The Kalman transition matrix model of the formant trajectories consists of a low order AR process of order 2–5; the process noise is the random process that drives the AR model of the formant trajectory; the observation noise is the noise-induced disturbance in formant track after the initial pre-cleaning of speech. It is worth noting that after pre-cleaning with MMSE, the effect of additive residual noise on the poles of a signal is rather complicated. However, in this work it is assumed that the residue noise after pre-cleaning manifests itself as an additive disturbance on the poles of the LP models. Kalman filters are described in Section 5.

3.6. Performance evaluation of formant tracking LP model

A formant-track percentage error measure, defined in Eq. (9), is used for the evaluation of the performance of the Kalman-based formant tracker described in Sections 3.1–3.5 for restoration of the formants of noisy speech. The reference formant tracks are obtained via a formant HMM trained on clean speech. The results of formant estimation with and without noise reduction and Kalman smoothing are shown in Fig. 5. The application of MMSE noise reduction results in significant improvement in reduction of formant tracking error. Further improvement in formant track estimation is obtained through application of Kalman filtering. Over 60% improvement in format track error through noise reduction has been achieved in the tracking the first formant, which is most affected by the noise. In less affected higher formants (F2–F5), the Kalman-based method recovers the formant track with an average of 15% improvement. Fig. 6 illustrates an example of formant recovery using a spectrogram of clean speech superimposed with the formant tracks of clean speech and the formant tracks, recovered from the noisy speech.

4. Estimation of harmonic noise model (HNM) of excitation

In this section, a harmonic plus noise model (HNM) of speech excitation, Eq. (3), is introduced. The HNM of noisy speech excitation is denoised with Kalman filters. Note that in the method proposed here an estimate of the noisy excitation is obtained through LP inverse filtering. Similarly, an estimate of the noise contaminating the speech excitation can be obtained by inverse filtering of the noise spectrum with the inverse LP model of noisy speech. The noisy excitation is then modeled by an HNM model and denoised by a Kaman filter which utilizes the temporal dependency of the trajectory of the excitation signal across speech frames.

The estimation and denoising of the parameters of the HNM of excitation, Eq. (3), of the LP model includes the followings steps:

(a) Voiced/unvoiced classification.
(b) Estimation, tracking and smoothing of the fundamental frequency and harmonic frequency tracks.
(c) Kalman filtering and smoothing of the noisy amplitudes of the harmonics.
(d) Denoising and estimation of the noise components of the speech excitation signal.

The estimation of HNM parameters is discussed next.

4.1. Fundamental frequency (pitch) estimation

Traditionally pitch is derived from the autocorrelation function as the inverse of the autocorrelation lag corresponding to the second largest peak of the autocorrelation function (Secrest and Doddington, 1983), note that the largest peak happens at the lag zero and corresponds to the signal energy. Since autocorrelation of a periodic signal is itself periodic, all the periodic peaks of the autocorrelation function can be usefully employed in the pitch estimation process as in Griffin's methods (Griffin and Lim, 1988). Other examples of pitch estimation include Friedman's work on development of a pseudo-maximum-likelihood pitch extraction method (Friedman, 1977) and Tucker' work on a voice activity detection based on periodicity detection (Tucker, 1992).

The pitch estimation method proposed in this work is an extension of the autocorrelation method to the frequency domain, it also incorporates signal to noise ratio at the harmonics. A pitch estimation error criterion over the speech harmonics is defined as

$$E(F_0) = E - F_0 \sum_{k=1}^{\max F} \sum_{l=kF_0-M}^{kF_0+M} W(l) \log |X(l)| \tag{10}$$

where $X(l)$ is the DFT of speech at discrete-frequency l, F_0 is a proposed value of the fundamental frequency (pitch) variable, $E = \sum_{l=0}^{\max F} \log |X(l)|$ and $2M + 1$ is a band of values about each harmonic frequency. The use of the logarithmic compression in Eq. (10) provides for a more balanced influence, on pitch estimation, between the high-energy low-frequency harmonics and the low-energy high-frequency harmonics. The weighting function $W(l)$ is a SNR-dependent Wiener-type weight given by

$$W(l) = \frac{\mathrm{SNR}(l)}{1 + \mathrm{SNR}(l)} \tag{11}$$

Fig. 7 shows an example of the variation of $E(F_0)$ curve with a range of pitch values. For each speech frame N pitch candidates are obtained as the N minimum values of $E(F_0)$ calculated on a grid of values of $F_{0\min} < F_0 < F_{0\max}$. The Viterbi algorithm is subsequently used to obtain the best pitch trajectory estimate through the given N candidates. Fig. 8 shows an example of speech and pitch and harmonic frequency tracks.

Fig. 9 shows a comparative illustration of the performance of the proposed pitch estimation method described here with Griffin's autocorrelation method, at different SNRs for car noise and train noise. It can be seen that as the SNR decreases, the autocorrelation-based method is less robust and degrades more than the proposed frequency-domain method. The proposed frequency method with SNR weighting provides improved performances in all cases we evaluated.

4.2. Estimation of harmonic amplitudes of excitation

The harmonic part of speech excitation signal of Eq. (3) is modeled as

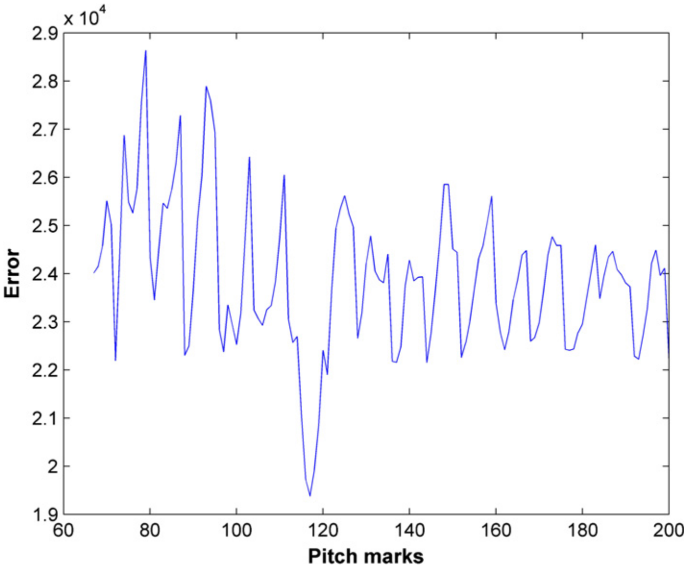

Fig. 7. An illustration of the variation of $E(F_0)$ curve with the proposed values of pitch frequency F_0.

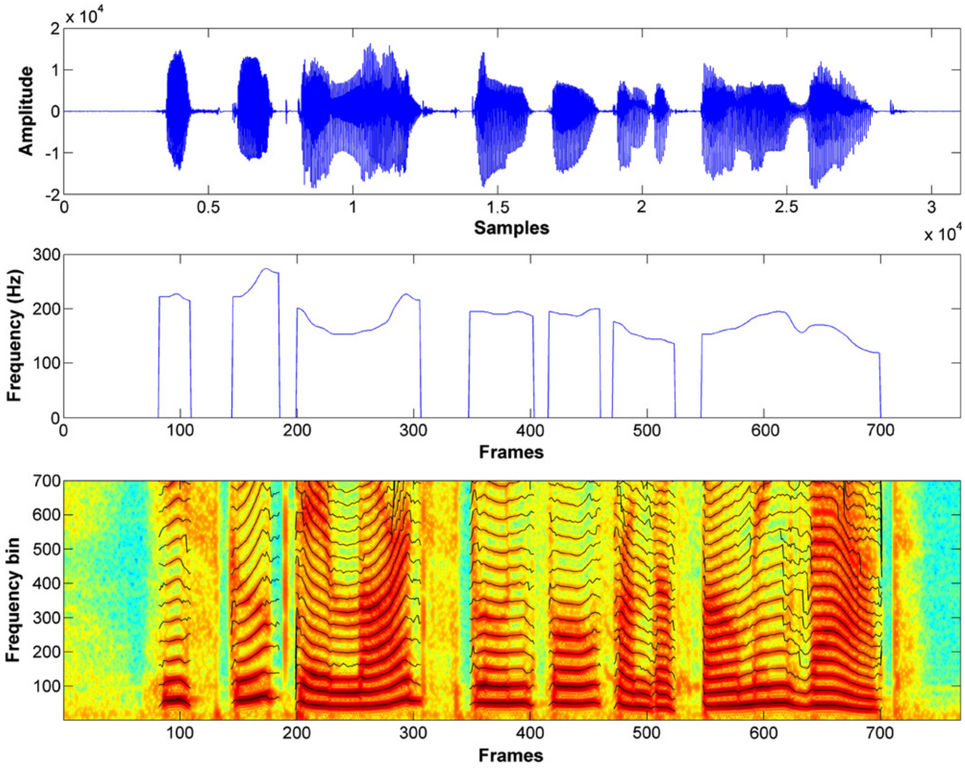

Fig. 8. An illustration of pitch tracks of a speech segment at sampling frequency of 8 kHz.

$$e_h(m) = \sum_{k=1}^{L(m)} A_k(m)\cos(2\pi(kF_0(m) + \Delta_k(m))m + \varphi_k(m)) + d(m) = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S} + d(m) \tag{12}$$

where $L(m)$ denotes the number of harmonics and $F_0(m)$ denotes the pitch, $\boldsymbol{A}$ and $\boldsymbol{S}$ are the harmonic amplitude vector and the harmonically related sinusoids vectors respectively and $d(m)$ is the noise contaminating the

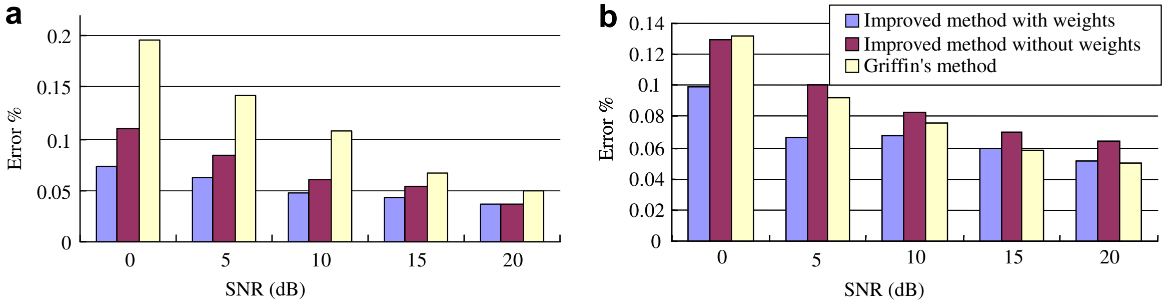

Fig. 9. Comparison of the performance of different pitch track methods for speech in: (a) train noise (b) car noise from 0 dB SNR to clean speech.

excitation. The maximum significant harmonic number $L(m)$ is obtained from the ability of the harmonic model to synthesis speech locally at the higher harmonics of the pitch (Seltzer et al., 2003). Note that the deviation of the harmonic frequencies, $\Delta_k(m)$, from the nominal value can be obtained from a search for the peak amplitudes about the nominal harmonic frequencies.

Given the harmonics frequencies, the harmonic amplitudes can be obtained either from searching for the peaks of the speech DFT spectrum or through a least square error estimation method as described in Stylianou (1996). To denoise the noisy excitation harmonics, to obtain estimates of the amplitudes of clean excitation harmonics, one can either apply a Kalman filter directly to the noisy excitation harmonics as in this paper, or alternatively the excitation can be denoised first with an eigen-based analysis method for estimation of sinusoids in noise (such as Esprit or Music, Vaseghi, 2006) and then the denoised harmonics can be smoothed using a set of Kalman filters. The Kalman filter method of denoising incorporates the temporal correlations of the successive samples of each harmonic in the denoising process.

4.3. Estimation of noise component of HNM

For unvoiced speech the excitation of the speech signal is a random noise-like process across the entire speech bandwidth. For voiced speech the excitation is mixture of harmonic and noise components.

Since after whitening of the signal with the inverse LP filter, the main effect of the background noise on the estimate of the noise component of the excitation is an increase in its variance; hence the noise part of the excitation may be de-noised by restoring its variance. Alternatively the noise part of the excitation signal may be replaced by a Gaussian random process of a similar variance to that of the noise component of speech excitation. In experiments on the use of FTLP-HNM model for speech synthesis we have obtained perceptually transparent results by replacing the noise part of the excitation to the LP model with a Gaussian noise of the appropriate variance, where the variance is estimated as the gain of the LP model of clean speech. Finally, the synthesized HNM of the excitation signal is obtained as the sum of the harmonic $\hat{e}_h(m)$ plus noise $\hat{e}_n(m)$ parts as

$$\hat{e}(m) = \hat{e}_h(m) + \hat{e}_n(m) \tag{13}$$

The synthesized excitation is combined with the formant-tracking LP model to reconstruct speech.

5. Kalman filtering of trajectories of formants and harmonics

Kalman filters (Kalman, 1960) are Bayesian models with a Markovian state transition matrix and Gaussian pdfs of process noise and observation noise. Kalman filters are used here to model, and smooth the trajectories of the LP–HNM namely; the formants, the pitch, the amplitudes of the harmonic tracks and the variance of unvoiced excitation. The Kalman filter formulation for all speech parameters is essentially the same, for this reason we describe the Kalman filters for estimation of formant tracks. Similar theory and equations hold for the smoothing of pitch, harmonics and the variance of unvoiced excitation.

Using the Kalman filter theory, the kth formant track sequence $\widehat{\boldsymbol{F}}_k(m) = [\widehat{F}_k(m), \widehat{F}_k(m-1), \widehat{F}_k(m-2), \ldots]$ is estimated from the trajectory of the formant track up to time $m-1$, $\widehat{\boldsymbol{F}}_k(m-1)$, and the current formant observation of the associated pole $p_k(m)$. The model used here to construct the Kalman state equation for the kth formant trajectory is an AR process defined as

$$\widehat{F}_k(m) = \sum_{i=1}^{P} c_{ki}\widehat{F}_k(m-i) + e_k(m) \tag{14}$$

where P is the model order (typically 4–5), and $e_k(m)$ is a zero mean Gaussian noise process; $p(e_k(m)) \sim N(0, Q_k)$. Note the AR model coefficients c_{ki} are the coefficients of the Kalman transition state matrix. The variance of the process noise Q_k is estimated recursively from the previous estimates of e_k. The use of a low order AR model follows from the observation that the trajectories of the formants are generally characterized by slow variation. The kth formant observation is obtained from the kth pole $p_k(m)$ modeled as

$$p_k(m) = \widehat{F}_k(m) + d_k(m) \tag{15}$$

where $d_k(m)$ is the noise in the estimates of the poles of the LP model due to residues left after pre-cleaning of speech spectral amplitude. The noise $d_k(m)$ is assumed to be a Gaussian zero mean process with variance of R_k, $p(R_k) \sim N(0, R_k)$. The variance of $d_k(m)$, R_k, is estimated recursively as the difference between the observed values of formants and the de-noised Kalman filtered estimates of formants. The algorithm for the discrete-time Kalman filter (Kalman, 1960) adapted for formant track estimation is as follows:

Time update (Predict) equations

$$\widehat{\boldsymbol{F}}_k(m|m-1) = \boldsymbol{C}\widehat{\boldsymbol{F}}_k(m-1) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_P & c_{P-1} & c_{P-2} & \cdots & c_1 \end{pmatrix} \begin{pmatrix} \widehat{F}_k(m-P) \\ \widehat{F}_k(m-P+1) \\ \widehat{F}_k(m-P+2) \\ \vdots \\ \widehat{F}_k(m-1) \end{pmatrix} \tag{16}$$

$$\boldsymbol{P}(m|m-1) = \boldsymbol{P}(m-1) + \boldsymbol{Q} \tag{17}$$

Measurement update equations

$$\boldsymbol{K}(m) = \boldsymbol{P}(m|m-1)(\boldsymbol{P}(m|m-1) + \boldsymbol{R})^{-1} \tag{18}$$

$$\widehat{\boldsymbol{F}}_k(m) = \widehat{\boldsymbol{F}}_k(m|m-1) + K(m)(\boldsymbol{p}_k(m) - \widehat{\boldsymbol{F}}_k(m|m-1)) \tag{19}$$

$$\boldsymbol{P}(m) = (\boldsymbol{I} - \boldsymbol{K}(m))\boldsymbol{P}(m|m-1) \tag{20}$$

where $\widehat{\boldsymbol{F}}_k(m|m-1)$ denotes a prediction of $\boldsymbol{F}_k(m)$from estimates of the formant track up to and including time m-1, $\boldsymbol{C}$ is the state transition matrix composed of the AR model coefficients; $\boldsymbol{P}(m)$ is the formant estimation error covariance matrix, $\boldsymbol{P}(m|m-1)$ is the formant prediction error covariance matrix, $\boldsymbol{K}(m)$ is the Kalman filter gain, $\boldsymbol{R}$ is the measurement noise covariance matrix, estimated from the variance of the differences between the noisy formant observation and estimated tracks. The covariance matrix $\boldsymbol{Q}$ of the process noise is obtained from the prediction error of formant tracks.

5.1. State-dependent Kalman filters

Kalman filter theory assumes that the signal and noise trajectories can be described by linear systems driven with random Gaussian excitation. A Kalman filter is unable to deal with the relatively sharp changes in the spectral characteristics of the signal process, for example when speech moves from a voiced to a non-voiced segment. State-dependent Kalman filters can be used to train and specialize Kalman filters to operate on different states of speech signal. In the simplest method used here, a two-state voiced/unvoiced classification of speech is used to employ two sets of Kalman filters; one set of Kalman filters for voiced speech and another set for unvoiced speech. It is worth noting that within voiced segments the formant trajectories are continuous and furthermore often there is some continuity between the formant trajectories of voiced speech on the two sides of an unvoiced segment.

6. Performance evaluation for speech enhancement

The databases used for the evaluation of the performance of formant trackers and speech enhancement systems are a subset of five male speakers and five female speakers from Wall Street Journal (WSJ). For each speaker, there are over 120 sentences. Speech signals are down sampled to 10 kHz from an original sampling rate of 16 kHz. The speech signal is segmented into overlapping frames of length 250 samples (25 ms) with an overlap of 150 samples (15 ms) between successive frames and hence a frame rate of 100 Hz.

6.1. Speech distortion measurements

The distortions measures used for evaluations of the speech enhancement method are Itakura-Saito Distance (ISD) and the harmonicity measure. The ISD measure (Deller et al., 1993) is defined as

$$\mathrm{ISD}_{12} = \frac{1}{L}\sum_{j=1}^{L}\frac{(\boldsymbol{a}_1(j)-\boldsymbol{a}_2(j))\times \boldsymbol{R}_1(j)\times(\boldsymbol{a}_1(j)-\boldsymbol{a}_2(j))^{\mathrm{T}}}{\boldsymbol{a}_1(j)\times \boldsymbol{R}_1(j)\times \boldsymbol{a}_1(j)^{\mathrm{T}}} \tag{21}$$

where $\boldsymbol{a}_1(j)$ and $\boldsymbol{a}_2(j)$ are the linear predication model coefficient vectors calculated from clean and transformed speech at frame j and $\boldsymbol{R}_1(j)$ is an autocorrelation matrix derived from the clean speech. The ISD criterion is a more balanced measure of the distance between an original clean speech signal and a distorted speech signal as speech frames with relatively large SNRs do not dominate the overall distance measure to the same extent as in the more conventional SNR measures.

An important aspect of the quality of voiced speech is its harmonicity defined as the ratio of the harmonic energy to noise energy of speech at and around each harmonics. The harmonicity parameter has been empolyed in a different form and context in speech synthesis, bandwidth extension and speech interpolation (Vaseghi et al., 2006). In general random noise degrades the harmonicity of speech. In this paper, for the purpose of measurement of the distortions of the harmonic structure of voiced speech, a harmonic contrast function is defined as the ratio of the peak signal power at harmonics to signal power in the troughs between the harmonics as

$$\mathrm{Harmonicity} = \frac{1}{NH\times N_{\mathrm{frames}}}\sum_{N_{\mathrm{frames}}}\sum_{k=1}^{NH}10\log_{10}\frac{P_k+P_{k+1}}{2P_{k,k+1}} \tag{22}$$

where P_k is the power at harmonic k, $P_{k,k+1}$ is the power at the trough between harmonics k and $k+1$, NH is the number of harmonics and N_{frames} is the number of speech frames.

Fig. 10 shows a comparison of ISD of noisy speech and speech restored with MMSE noise reduction method and speech restored with LP–HNM method. The FTLP-HNM method performs better than MMSE in improving the ISD of noisy speech relative to the clean speech.

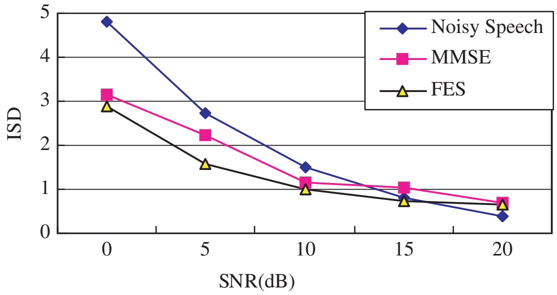

Fig. 10. Comparison of ISD of noisy speech in train noise pre-cleaned with MMSE and improved with FTLP-HNM system (FES) at SNR = 0, 5, 10, 15 dB.

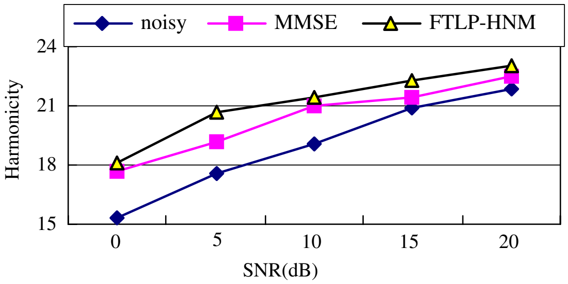

Fig. 11. Comparison of the harmonicity of MMSE and FTLP-HNM systems on noisy speech (train noise) at different SNRs.

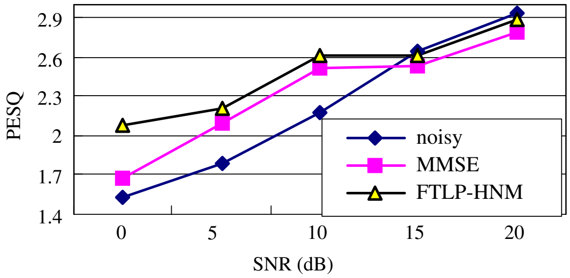

Fig. 12. Performance of MMSE and FTLP-HNM on noisy speech (train noise) at different SNRs.

Fig. 11 shows the improvement in the harmonicity that can be obtained with FTLP-HNM model compared to that of noisy speech and speech enhanced with MMSE. The figure shows that a consistent improvement in harmonicity is obtained from LP–HNM model denoised with Kalman filters.

Fig. 12 shows the improvement in the perceptual evaluation of speech quality (PESQ) (PESQ) of restored speech resulting from application of FTLP-HNM model in comparison to noisy speech and speech cleaned with the MMSE method. It is significant that in all cases evaluated the FTLP-HNM model delivers improved results.

7. Conclusion

This paper presented a parameter-tracking LP model combined with a harmonic and noise model of the excitation for enhancement of noisy speech. The proposed method utilizes the spectral-temporal structures of speech. An important feature of the proposed method is the tracking of the dominant energy contours of the spectral envelop and the harmonics of the excitation of speech using Viterbi trackers followed by Kalman filters. The estimates of clean LP model are smoothed with Kalman filters and the noisy excitation is denoised with Kalman filters. Evaluations of the de-noising system shows that it delivers improved results compared to MMSE method with significantly less artifacts such as ‘musical noise’ also known as ‘musical tones’.

The speech enhancement method described here can be implemented for real-time and non-real-time applications. The total delay is the sum of the segment delay plus the processing delay. The segment delay is the frame duration which is usualy 20 ms. The processing delay depends on the order (number) of the past samples used in Kalman filter and the Viterbi sequence estimator and also on whether tracking is based on use of HMMs or on the use of a simple MMSE distance of the current sample value of a paramter trajectory from its past values. Note that in non-real-time applications the future values of trajectories of various parameters can be usefully employed.

The method is currently extended to restoration of speech signals where significant parts of the speech spectrum are missing or lost to noise.

Acknowledgement

We thank the UK's EPSRC for funding project No. GR/S30238/01.

References

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustic Speech and Signal Processing ASSP-27, 113–120.

Chen, A., Vaseghi, S., McCourt, P., 2000. State based sub-band LP wiener filters for speech enhancement in car environments. Proceedings of ICASSP I, 213–216.

Deller Jr., J.R., Proakis, J.G., Hansen, J.H.H., 1993. Discrete-Time Processing of Speech Signals. Macmillan, New York.

Ephraim, Malah D., 1985. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Transactions on Acoustic Speech and Signal Processing ASSP-33, 443–445.

Ephraim, Y., 1992. Statistical-model based speech enhancement systems. Proceedings of the IEEE 80 (10), 1526–1554.

Ephraim, Y., Van Trees, Harry L., 1995. A signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing 3 (4), 251–266.

Ephraim, Y., Malah, D., Juang, B.-H., 1989. On the application of hidden Markov models for enhancing noisy speech. IEEE Transactions on Acoustic Speech and Signal Processing ASSP-37, 1846–1856.

Friedman, D., 1977. Pseudo-maximum-likelihood speech pitch extraction. Transaction on ASSP, 213–221.

Griffin, D.W., Lim, J.S., 1988. Multiband-excitation vocoder. IEEE Transactions on Acoustic Speech and Signal Processing ASSP-36 (2), 236–243.

Hansen, J.H.L., Clements, M.A., 1987. Iterative speech enhancement with spectral constraints. Proceedings of ICASSP, 189–192.

Kalman, R., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME, Journal of Basing Engineering 82, 34–35.

Kim, C., Sung, W., 2001. Vowel pronunciation accuracy checking system based on phoneme segmentation and formants extraction. In: Proceedings of the International Conference on Speech Processing, Daejeon, Korea. pp. 447–452.

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Transactions on Acoustic Speech and Signal Processing ASSP-26 (3), 197–210.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proceedings of IEEE 67, 1586–1604, Dec.

PESQ available at http://www.pesq.org/.

Rentzos, D., Vaseghi, S., Yan, Q., Ho, C., Turajlic, E., 2003. Probability models of formant parameters for voice conversion. Proceedings of Eurospeech, 2405–2408.

Sameti, H., Sheikhzadeh, H., Deng, L., Brennan, R.L., 1998. HMM-based strategies for enhancement of speech signals embedded in non-stationary noise. IEEE Transactions on Speech and Audio Processing 6 (5), 445–455.

Secrest, B., Doddington, G., 1983. An integrated pitch tracking algorithm for speech systems. Proceedings of ICASSP, 1352–1355.

Seltzer, M.L., Droppo, J., Acero, A., 2003. A harmonic-model-based front end for robust speech recognition. Proceedings of Eurospeech, 1277–1280.

Stylianou, Y., 1996. A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech. IEEE Nordic Signal Processing Symposium.

Stylianou, Y., 1996. A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech. IEEE Noric Signal Processing Symposium (September).

Tucker, R., 1992. Voice activity detection using a periodicity measure. IEE Proceedings Communications, Speech and Vision 139 (4), 377–380. August.

Turunen, J., Vlaj, D., 2001. A study of speech coding parameters in speech recognition. Proceedings of Eurospeech, 2363–2366.

Vaseghi, S., 2006. Advanced Digital Signal Processing and Noise Reduction, third ed. Wiley, New York.

Vaseghi, S., Zavarehei, E., Yan, Q., 2006. Speech bandwidth extension: extrapolations of spectral envelop and harmonicity quality of excitation. In: ICASSP 2006 Proceedings Toulouse.

Weber, K., Bengio, S., Bourlard, H., 2001. HMM2-extraction of formant structures and their use for robust ASR. Proceedings of Eurospeech, 607–610.

Yan, Q., Zavarehei, E., Vaseghi, S., Rentzos, D., 2004. A formant tracking LP model for speech processing in car/train noise. Proceedings of ICSLP, 734–737.