KALMAN FILTER WITH LINEAR PREDICTOR AND HARMONIC NOISE MODELS FOR NOISY SPEECH ENHANCEMENT

*Qin Yan   Saeed Vaseghi   Esfandiar Zavarehei   Ben Milner**

School of Engineering and Design, Brunel University, London , *School of Computing Sciences, University of East Anglia, Norwich, UK,
{Qin.Yan, Saeed.Vaseghi, Esfandiar.Zavarehei}@brunel.ac.uk, {bpm*}@cmp.uea.ac.uk

ABSTRACT

This paper presents a method for noisy speech enhancement based on integration of a formant-tracking linear prediction (FTLP) model of spectral envelope and a harmonic noise model (HNM) of the excitation of speech. The time-varying trajectories of the parameters of the LP and HNM models are tracked with Viterbi classifiers and smoothed with Kalman filters. A frequency domain pitch estimation is proposed, that searches for the peak SNRs at the harmonics. The LP-HNM model is used to deconstruct noisy speech, de-noise its LP and HNM models and then reconstitute cleaned speech. Experimental evaluations show the performance gains resulting from the formant tracking, harmonic extraction and noise reduction stages.

1. INTRODUCTION

Linear prediction (LP) and harmonic noise models (HNM) [1] are the two main methods for modeling speech waveforms. LP and HNM offer complementary advantages; LP model provides a good fit for the spectral envelope whereas HNM is good at modeling the fine details of the harmonic plus noise structure of speech excitation.

The motivation for the proposed integration, of LP and HNM, is to model and untilise the *spectral-temporal* trajectories of the dominant parameters of speech. For noisy speech processing this is a different approach to spectral amplitude estimation methods [2] which generally model each individual spectral sample in isolation without fully utilizing the wider spectral-temporal structures that may be used to good effect in the de-noising process to obtain improved results.

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 concentration contours also exist for unvoiced speech at relatively higher frequencies

In this paper HNMs are used to model the trajectories of the excitation of LP model. Harmonic noise models (HNM) are an established method particularly in speech and music coding and text to speech synthesis [1]. The main issues in HNM are voiced/unvoiced classification and the estimations of the fundamental frequency (pitch) value, the harmonic amplitudes and the noise component of speech excitation.

Previous work related to the FTLP-HNM includes the use of Kalman filters for formant estimation [3] and the use of HNM for speech enhancement [4]. The distinctive contribution of this paper is the integration of LP and HNM models with Viterbi classifiers and Kalman filters for tracking and de-noising the trajectories of the model parameters for enhancement of noisy speech.

2. AN OVERVIEW OF FORMANT-TRACKING LP MODEL WITH HNM EXCITATION

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

(1) A pre-cleaning module for de-noising speech prior to the estimation of the LP model and formant parameters.

(2) A formant-tracking LP model estimation incorporating Viterbi trackers and Kalman smoothers.

(3) A pitch extraction method incorporating Viterbi trackers and Kalman smoothers.

(4) A harmonic noise model estimation method using Kalman filters for noise reduction and smoothing.

The LP model of speech $X(z,m)$ may be expressed as

$$X(z,m) = E(z,m)V(z,m) \qquad (1)$$

where $E(z,m)$ is the z-transform of the excitation signal and $V(z,m)$ is the z-transform of a LP model of the combined effect of the

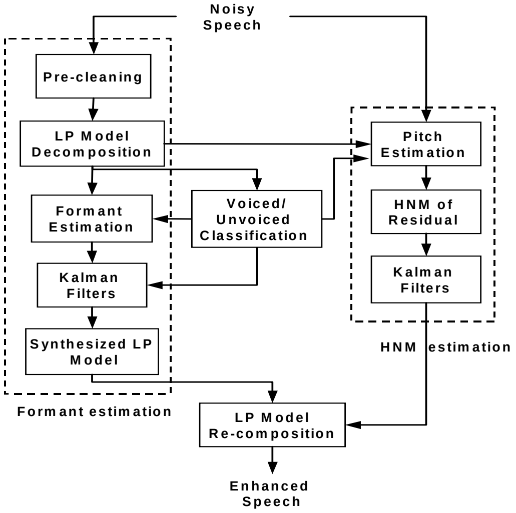

Figure 1 – Overview of the FTLP-HNM model for enhancement of noisy speech.

vocal tract, the glottal pulse and the lip radiation. $V(z,m)$ can be expressed as a cascade of a set of second order model of resonances and a first order model of the spectral slope as

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

where $r_k(m)$ and $\varphi_k(m)$ 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.

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

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

where $F_0(m)$ is the time-varying fundamental frequency of excitation, $A_k(m)$ are excitation harmonics, $\varphi_k(m)$ is the phase and $v(m)$ is the noise part of the excitation. In the following the estimation of the parameters of FTLP and the HNM models are described.

3. FORMANT ESTIMATION FROM NOISY SPEECH

In this section a robust formant-tracking LP model is introduced composed of pre-cleaning of speech spectrum followed by formant track estimation and Kalman smoothing of formant tracks.

3.1 Initial-Cleaning of Noisy Speech

Before formant estimation, noisy speech spectrum is pre-cleaned using the MMSE spectral amplitude estimation method [5]. 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 using the Levinson-Durbin method. A formant tracker is then used to process the poles of the LP model and obtain an improved estimate of the LP model parameters as described next.

3.2 Formant Tracking

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

$$\mathbf{v}_k = [F_k, B_k, M_k, \Delta F_k, \Delta B_k, \Delta M_k] \qquad k=1, \ldots, N \quad (4)$$

where the number of formants is typically set to N=5. Velocity derivatives, denoted by Δ, are computed as the slopes of the features over time. The probability distributions of formants can be modeled by Gaussian mixture model (GMM) or HMMs as described in detail in [6]. A Viterbi classifier is used to classify and track the poles of the LP model associated with different formants. Klaman filters, described in section 5, are subsequently employed to smooth formant trajectories [3]. Note that instead of formants one can equivalently employ the line spectral frequencies.

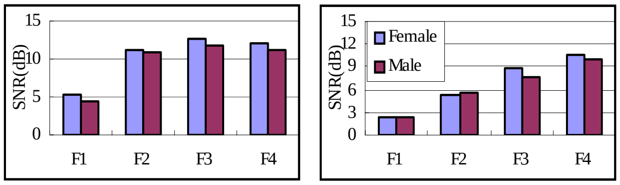

Figure 2– Variation of speech SNR at different formants in (left) car noise (right) train noise at average SNR=0 dB**.**

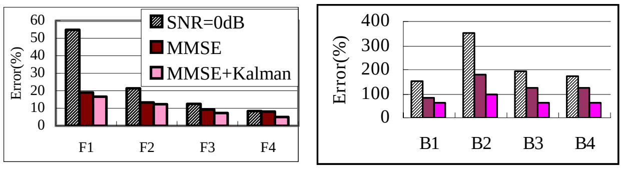

Figure 3- Average % error of formant tracks (frequency F_k and bandwidth B_k) in train noise and cleaned speech using MMSE and Kalman filters, the results were averaged over five males.

The speech database used to investigate the effect of noise on formants is the Wall Street Journal. The speech is degraded by car noise or train noise with an average SNR in the range from 0 to 20 dB. To quantify the contamination of formants by noise a local formant signal to noise ratio measure (FSNR) [3] is defined as

$$FSNR(k) = 10\log\left[\sum_{l\in(F_k\pm B_k/2)} X_l^2 \Big/ \sum_{l\in(F_k\pm B_k/2)} N_l^2\right] \quad (5)$$

where X_l and N_l are the magnitude spectra of speech and noise respectively and F_k and B_k are the frequency and bandwidth of the k^{th} formant. Figure 2 displays the FSNRs of noisy speech in moving car and train environments. It is evident that the FSNRs are higher than the overall SNR.

To quantify the effects of the noise on formant estimation, an average formant track error measure, defined as

$$E_k = \frac{1}{L}\sum_{m=1}^{L}\left[\left|F_k(m)-\hat{F}_k(m)\right|\Big/F_k(m)\right]\times 100\% \quad k = 1,\ldots,N \quad (6)$$

where $F_k(m)$ and $\hat{F}_k(m)$ 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.

Figure 3 shows the improvement in formant estimation resulting from pre-cleaning followed by Viterbi classifier and Kalman filters. The reference formant tracks are obtained from HMMs of formants of clean speech [6]. The application of MMSE noise suppression results in significant reduction of formant tracking error. Further improvement is obtained through application of Kalman filtering. Over 60% improvement in format track error through noise reduction is achieved in the tracking of 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.

4. HARMONIC NOISE MODEL OF SPEECH EXCITATION

The estimation of the parameters of the harmonic plus noise model of the excitation includes the followings steps:

(a) Voiced/Unvoiced classification.

(b) Estimation and smoothing of the fundamental frequency and harmonic tracks.

(c) Estimation and smoothing of the amplitudes of harmonics.

(d) Estimation of the noise component of the excitation.

The estimation of HNM parameters is discussed next.

4.1 Fundamental Frequency Estimation

Traditionally pitch is derived as the inverse of the time τ corresponding to the second largest peak of the autocorrelation of speech. Since autocorrelation of a periodic signal is itself periodic, all the periodic peaks of the autocorrelation can be used in the pitch estimation process [7]. The proposed pitch estimation method is an extension of the autocorrelation-based method [7] to frequency domain. A pitch estimation error is defined as

$$E(F_0) = E - F_0 \sum_{k=1}^{MaxF} \sum_{l=kF_0-M}^{kF_0+M} W(l)\log|X(l)| \qquad (7)$$

where $X(l)$ is the DFT of speech, F_0 is a proposed value of the fundamental frequency (pitch) variable, E is sum of log spectral energy, and $2M+1$ is a band of values about each harmonic frequency. The weighting function $W(l)$ is a SNR-dependent Wiener-type weight. Figures 4 provides a comparative illustration of the performance of the proposed pitch estimation method with Griffin's method [7], at different SNRs for car noise and train noise. It can be seen that the proposed frequency method with SNR weighting provides improved performances in all cases evaluated.

4.2 Harmonic Amplitudes Estimation

The harmonics of speech excitation is modeled as

$$e_h(m) = \sum_{k=1}^{L(m)} A_k(m)\cos\left(2\pi k F_0(m)m + \varphi_k(m)\right) = \boldsymbol{A}^T\boldsymbol{S} \qquad (8)$$

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 vector respectively. Given the harmonics frequencies, the amplitudes $\boldsymbol{A}$ can be obtained either from searching for the peaks of the speech DFT spectrum or through a least square error estimation. The maximum significant harmonic number is obtained from the ability of the harmonic model to synthesis speech locally at the higher harmonics of the pitch [8].

The estimate of the amplitudes of clean excitation harmonics is obtained from a set of Kalman filters one for each harmonic. The Kalman filter is the preferred method here as it models the trajectory of the successive samples of each harmonic.

4.3 Estimation of Noise Component of HNM

For unvoiced speech the excitation is noise-like across the entire speech bandwidth. For voiced speech the excitation is noise-like above some variable maximum harmonic frequency.

The main effect of the background noises on the estimate of the excitation of LP model is an increase its variance. We have obtained perceptually good results by replacing the noise part of the excitation to LP model with a Gaussian noise with the appropriate variance estimated as the difference between the variance of the noisy signal and that of the noise. Finally the synthetic HNM of excitation signal is obtained as

$$\hat{e}(m) = e_h(m) + e_n(m) \qquad (9)$$

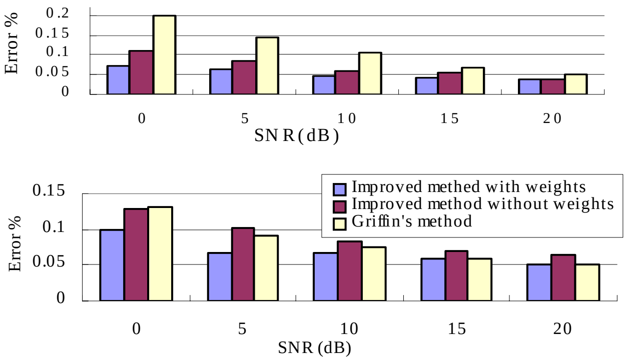

Figure 4 - Comparison of different pitch track methods for speech in train noise (top) car noise (bottom) from 0dB SNR to clean.

5 KALMAN SMOOTHING OF TRAJECTORIES OF FORMANTS AND HARMONICS

The Kalman filter equations for all the parameters of speech are essentially the same, for this reason we describe the kalman smoothing of formant tracks. The formant trajectory is modeled by an AR process as

$$\hat{F}_k(m) = \sum_{i=1}^{P} c_{ki}\hat{F}_k(m-i) + e_k(m) \qquad (10)$$

where c_{ki} are the coefficients of a low order (3 to 5)AR model of the k^{th} formant track and $e_k(m)=N(0,Q_k)$ is a zero mean Gaussian random process. The variance of $e_k(m)$, Q_k is estimated from the previous estimates of e_k. The algorithm for Kalman filter [9] adapted for formant track estimation is as follows.

Time updates (Prediction) equations

$$\hat{\boldsymbol{F}}_k(m \mid m-1) = \boldsymbol{C}\hat{\boldsymbol{F}}_k(m-1) \qquad (11)$$

$$\boldsymbol{P}(m \mid m-1) = \boldsymbol{P}(m-1) + \boldsymbol{Q} \qquad (12)$$

Measurement updates (Estimation) equations

$$\boldsymbol{K}(m) = \boldsymbol{P}(m \mid m-1)\left(\boldsymbol{P}(m \mid m-1) + \boldsymbol{R}\right)^{-1} \qquad (13)$$

$$\hat{\boldsymbol{F}}_k(m) = \hat{\boldsymbol{F}}_k(m \mid m-1) + \boldsymbol{K}(m)\left(\boldsymbol{p}_k(m) - \hat{\boldsymbol{F}}_k(m \mid m-1)\right) \qquad (14)$$

$$\boldsymbol{P}(m) = \left(\boldsymbol{I} - \boldsymbol{K}(m)\right)\boldsymbol{P}(m \mid m-1) \qquad (15)$$

where $\hat{F}_k(m \mid m-1)$ denotes a prediction of $F_k(m)$ from estimates of the formant track up to time m-1, $\boldsymbol{P}(m)$ is the formant estimation error covariance matrix, $P(m|m\text{-}1)$ 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.

Kalman theory assumes the signal and noise can be described by linear systems with random Gaussian excitation. Kalman filter is unable to deal with the relatively sharp changes in the signal

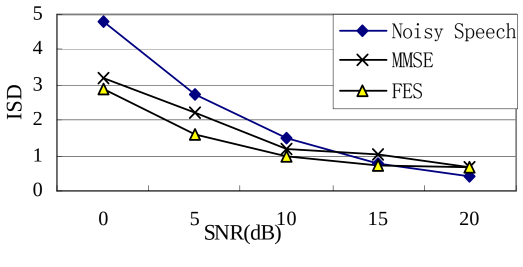

Figure 5 - Comparison of ISD of noisy speech in train noise pre-cleaned with MMSE and improved with formant-base enhancement system (FES) at SNR = 0, 5, 10, 15 dB.

process, for example when speech changes from a voiced to a non-voiced segment. However, state-dependent Kalman filters can be used to solve this problem. For example a two state voiced/unvoiced classification of speech can be used to employ two separate sets of Kalman filters; one set of Kalman filters for voiced speech and another set for unvoiced speech. In HMM-based speech models in each state of HMM the signal trajectory can be modelled a Kalman filter.

6. PERFORMANCE EVALUATION OF SPEECH ENHANCEMENT

The databases used for the evaluation of the performance of the 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 signal is 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.

The following distortion measures are used. The Itakura-Saito Distance (ISD) measure [10] is defined as

$$ISD_{12} = \frac{1}{L}\sum_{\boldsymbol{j}=1}^{\boldsymbol{L}} \frac{\left(\boldsymbol{a}_1(j) - \boldsymbol{a}_2(j)\right) \times \boldsymbol{R}_1(j) \times \left(\boldsymbol{a}_1(j) - \boldsymbol{a}_2(j)\right)^{\boldsymbol{T}}}{\boldsymbol{a}_1(j) \times \boldsymbol{R}_1(j) \times \boldsymbol{a}_1(j)^{\boldsymbol{T}}} \qquad (16)$$

where $\boldsymbol{a}_1(j)$ and $\boldsymbol{a}_2(j)$ are the LP coefficient vectors calculated from clean and processed speech at frame j and $\boldsymbol{R}_1(j)$ is an autocorrelation matrix of clean speech.

To measure the distortions of the harmonic structure of speech, a harmonic contrast function is defined as

$$Harmonicity = \frac{1}{NH \times N_{frames}} \sum_{N_{frames}} \sum_{k=1}^{NH} 10 \log \frac{P_k + P_{k+1}}{2P_{k,k+1}} \qquad (17)$$

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.

Figure 5 shows the improvement in ISD measure compared with MMS system. It is evident that the new speech processing system achieves a better ISD score. Figure 6 shows the significant improvement in the harmonicity measure resulting from FTLP-HNM model. Figure 7 illustrates the results of Perceptual Evaluation of Speech quality (PESQ) of noisy speech and speech restored with MMSE and FTLP-HNM methods. It s evident that in all respects FTLP-HNM method achieves improved results. Examples of contaminated and restored speech files are also available in http://dea.brunel.ac.uk/cmsp/florence_nighingale.htm

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. Evaluations of the de-noising system shows that it delivers improved results compared to MMSE method with significantly less artifacts such as musical noise. The method is currently extended to restoration of speech signals where significant parts of the speech spectrum are missing or lost to noise.

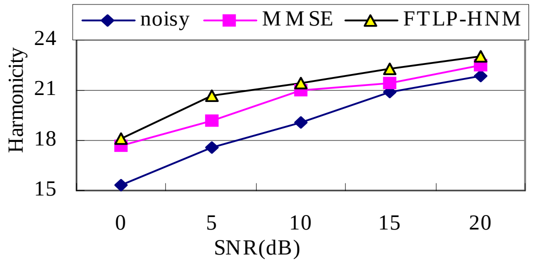

Figure 6-Comparison of harmonicity of MMSE and FTLP-HNM systems on train noisy speech at different SNRs.

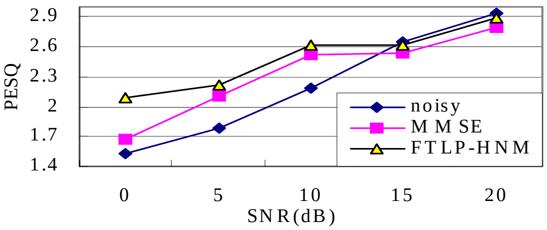

Figure 7-Performance of MMSE and FTLP-HNM on train noisy speech at different SNRs.

ACKNOWLEDGEMENT

We wish to thank the UK's EPSRC for funding this project.

REFRENCES

[1] Stylianou Y. "*A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech"* IEEE Nordic Signal Processing Symp., (1996).

[2] Vaseghi S., "*Advanced Digital Signal Processing and Noise Reduction"*, John Wiley, 3nd Ed. (2005).

[3] Yan Q, Vaseghi S., Zavarehei E., Milner B., "*Formant-Tracking Linear Prediction Model For Speech Processing In Noisy Environment*" Eurospeech (2005)

[4] Palpous C., Marro C. Scalart P. "*Speech Enhancement Using Harmonic Regeneration*" Proc. ICASSP pp.157-160(2005)

[5] Ephraim, Malah D., "*Speech Enhancement Using A Minimum Mean Square Error Log-Spectral Amplitude Estimator*" IEEE Trans. ASSP, Vol. -33, pp.443-445 (1985)

[6] Yan Q, Vaseghi S. "*Analysis, Modelling and Synthesis of formants of British, American and Australian Accents*" ICASSP, (2003).

[7] Griffin D. W., Lim J.S. "*Multiband-excitation vocode*r" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-36(2) pp.236-243 (1988).

[8] Stylianou Y., "*A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech*", IEEE Noric Signal Processing Symp. Sept (1996)

[9] Kalman R., "*A New Approach to Linear Filtering and Prediction Problems*", Transactions of the ASME, Journal of Basing Engineering, vol. 82, pp. 34-35 (1960)

[10] Deller J.R., Jr., Proakis, J.G., Hansen, J.H.H., *Discrete-Time Processing of Speech Signals*, New York: Macmillan Publishing Company (1993).