RESTORATION OF NOISY AND BAND LIMITED ARCHIVED SPEECH RECORDS WITH LINEAR PREDICTOR AND HARMONIC NOISE MODELS

*Qin Yan Saeed Vasegh*i Esfandiar Zavarehei Ben Milner***
School of Computer and Information Engineering, Hohai University, Nanjing, P.R.China
*School of Engineering and Design, Brunel University, London ,
**School of Computing Sciences, University of East Anglia, Norwich, UK,
yanqin@ieee.org *{Saeed.Vaseghi, Esfandiar.Zavarehei}@brunel.ac.uk, **bpm@cmp.uea.ac.uk

ABSTRACT

A method is presented for restoration of noisy bandlimited archived speech records. Speech is modeled with a formant-tracking linear prediction (FTLP) model of the spectral envelope and a harmonic noise model (HNM) of the excitation. The time-varying trajectories of the parameters of the LP and HNM models are tracked with Viterbi classifiers and denoised with Kalman filters. A frequency domain pitch estimation is proposed, which 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 the cleaned speech. The missing spectrum at lower and higher frequency bands are reconstructed through spectral extrapolation of the LP-HNM model. Comparative evaluations show the performance gains obtained from the proposed method.

***Index Terms*:** harmonic, linear prediction, Kalman filter

1. INTRODUCTION

This paper describes a speech enhancement method for restoration of archived speech records. The restoration method is based on using Kalman filters to smooth a precleaned linear prediction (LP) model of the spectral envelop and to denoise a harmonic noise model (HNM) of noisy excitation of speech [1].

The motivation for integration of LP and HNM models is to utilize the spectral-temporal trajectories of the prominent energy contours of speech. For noisy speech processing this is a different approach to spectral amplitude estimation methods [2] which generally model each spectral sample in isolation, without fully utilizing the wider spectral-temporal structures.

The proposed 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 harmonic noise models (HNM) are used to model the trajectories of the excitation of LP model [1]. Previous work related to the LP-HNM includes the use of Kalman filters for formant estimation [3] and the use of HNM for speech enhancement [4]. The distinctive contributions of this paper are:

(1) Integration of an LP-HNM model with Viterbi classifiers and Kalman filters for tracking and de-noising the trajectories of the speech model parameters.

(2) Extrapolation Spectrum of the denoised LP-HNM model in order to reconstruct the missing frequency bands of speech.

2. OVERVIEW OF SPEECH ENHANCEMENT METHOD

The proposed speech enhancement method is illustrated in Figure 1 and consists of the following sections:

(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 formants and poles of the LP model.

(3) A pitch extraction method incorporating Viterbi decoders and Kalman filters for pitch smoothing.

(4) A method for estimation of a harmonic noise model of clean excitation with Kalman filters used for modeling and denoising the temporal trajectory of noisy excitation.

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

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

where $E(z,m)$ is the z-transform of the excitation signal and $V(z,m)$, the z-transform of a LP model of the spectral envelop of speech can be expressed as

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

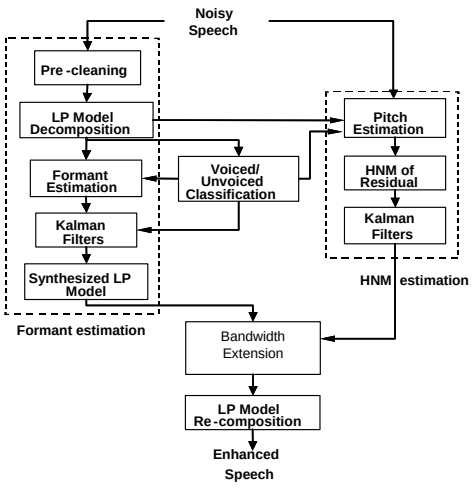

Figure 1 – An illustration of the proposed system for restoration of noisy bandlimited speech signals.

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

$$E(f,m)=\sum_{k=1}^{L(m)} A_k(m)G\big(f-kF_0(m)+\Delta_k(f,m)\big)+V(f,m) \quad (3)$$

where f is the frequency variable, $L(m)$ denotes the number of harmonics, $F_0(m)$ is the time-varying fundamental frequency, $\Delta_k(f,m)$ is the deviation of the k^{th} harmonic from the nominal value of kF_0, $A_k(m)$ are the complex amplitudes of excitation harmonics, $G(f)$ is a Gaussian-shaped model of harmonic shape, and $V(f,m)$ is the noise part of the excitation. The harmonic shape function $G(f)$ has a frequency support equal to F_0 and is selected to model the shape of the harmonics of speech in the frequency domain.

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 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 to obtain an improved estimate 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, $\boldsymbol{v}_k$ comprising the frequency, F_k, bandwidth, B_k and magnitude, M_k, of the resonance at formants together with their temporal slopes as

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

The number of formants is typically set to N=5. The probability distributions of formants can be modeled by Gaussian mixture model (GMM) as described in detail in [6]. A Viterbi classifier is used to classify and track the poles of the LP model associated with formants. Kalman filters, described in section 5, are employed to model and smooth the formant trajectories [3]. Note that instead of formants one can employ line spectral frequencies.

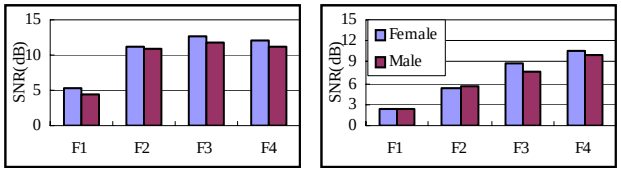

Figure 2– Variation of speech SNR at different formants in: (left) car noise and (right) train noise at average SNR=0 dB**.**

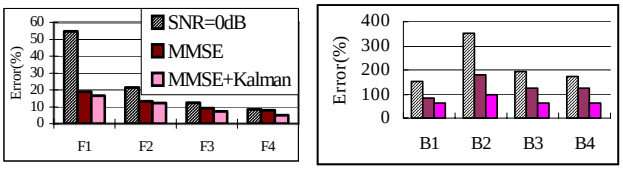

Figure 3- Average % error of formant tracks (frequency F_k and bandwidth B_k) in train noise and cleaned speech using MMSE and Kalman filters, the results were averaged over five males.

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 and F_k and B_k are the frequency and bandwidth of the k^{th} formant respectively. Figure 2 displays the FSNRs of noisy speech in moving car and train environments. It is evident that the FSNRs are substantially higher than the average SNR.

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

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

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. The reference formant tracks are obtained from HMMs of formants of clean speech [6]. The MMSE noise suppression results in significant reduction of formant tracking errors. Further improvement is obtained through Kalman filtering. Over 60% improvement in format track error is achieved for 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 EXCITATION

The section describes the denoising and estimation of the parameters of the harmonic plus noise model of the excitation.

4.1 Fundamental Frequency Estimation

Traditionally pitch is derived as the inverse of the time τ corresponding to the period of the autocorrelation of speech [7]. The pitch estimation error criterion used in this paper is defined as

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

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. Figure 4 provides a comparative illustration of the performance of the proposed pitch estimation method with Griffin's method [7] 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

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 and simultaneously denoises speech.

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.

5 KALMAN SMOOTHING OF LP-HNM TRACKS

The Kalman filter equations for all speech parameters 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

$$\hat{F}_k(m) = \sum_{i=1}^{P} c_{ki}\hat{F}_k(m-i) + e_k(m) \tag{8}$$

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 process. The variance of $e_k(m)$, Q_k is estimated from the previous estimates of e_k. The algorithm for Kalman filter [9] is as follows.

Time updates (Prediction) equations

$$\hat{\boldsymbol{F}}_k(m \mid m-1) = \boldsymbol{C}\hat{\boldsymbol{F}}_k(m-1) \tag{9}$$

$$\boldsymbol{P}(m \mid m-1) = \boldsymbol{P}(m-1) + \boldsymbol{Q} \tag{10}$$

Measurement updates (Estimation) equations

$$\boldsymbol{K}(m) = \boldsymbol{P}(m \mid m-1)\left(\boldsymbol{P}(m \mid m-1) + \boldsymbol{R}\right)^{-1} \tag{11}$$

$$\hat{\boldsymbol{F}}_k(m) = \hat{\boldsymbol{F}}_k(m \mid m-1) + \boldsymbol{K}(m)\left(\boldsymbol{p}_k(m) - \hat{\boldsymbol{F}}_k(m \mid m-1)\right) \tag{12}$$

$$\boldsymbol{P}(m) = \left(\boldsymbol{I} - \boldsymbol{K}(m)\right)\boldsymbol{P}(m \mid m-1) \tag{13}$$

where $\hat{F}_k(m \mid m-1)$ denotes a prediction of $F_k(m)$, $\boldsymbol{P}(m)$ is the estimation error covariance matrix, $P(m|m\text{-}1)$ is the 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. Kalman filter is unable to deal with sharp changes in the signal 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.

6 RESTORATION OF MISSING SPECTRUM

Old heritage archived speech records often have a limited frequency bandwidth between 400 to 3000 Hz or worse. Using the LP-HNM model the spectral envelop and the excitation models are restored (extrapolated) separately and then combined.

The restoration of the spectral envelop can be achieved using one of two alternative methods: (1) The extrapolation of the frequency response of the LP model to higher frequencies through upsampling the model's frequency response. This would have the effect of the extending the response of the poles, identified in the available bandwidth to higher and lower frequencies, (2) extraction of the missing spectrum from a codebook trained on joint features extracted from the narrowband and wideband speech. In this work we use the first method for its simplicity. Restoration of the harmonics of excitation is achieved through the following formula:

$$F_h(k,m) = \begin{cases} F_h(k-1,m) + F_0(m) & k > 1 \\ F_h(k+1,m) - F_0(m) & k = 1 \end{cases} \tag{14}$$

where $F_h(k,m)$ is the frequency of ith harmonic, $F_0(m)$ is the fundamental frequency at time *m*. The corresponding amplitude of harmonics are obtained from the excitation spectrum at $F_h(k,m)$. Note $F_h(k,m)$ is estimated from DFT spectrum of the whole speech while amplitude of harmonics are from the excitation spectrum. The phases of the missing harmonics are extrapolated linearly from the unwrapped phase of the available harmonics. For extrapolation of unvoiced excitation we use a Gaussian noise which is then filtered with the extended bandwidth LP model.

7. PERFORMANCE EVALUATION

The databases used for the evaluation of the performance of the speech enhancement systems are a subset of five male and five female speakers from Wall Street Journal (WSJ). For each speaker, there are over 120 sentences. The speech 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 for performance assessment. The Itakura-Saito Distance measure(ISD) [10] is defined as

$$ISD_{12} = \frac{1}{L}\sum_{j=1}^{L} \frac{\left(\boldsymbol{a}_1(j) - \boldsymbol{a}_2(j)\right) \times \boldsymbol{R}_1(j) \times \left(\boldsymbol{a}_1(j) - \boldsymbol{a}_2(j)\right)^{\boldsymbol{T}}}{\boldsymbol{a}_1(j) \times \boldsymbol{R}_1(j) \times \boldsymbol{a}_1(j)^{\boldsymbol{T}}} \tag{15}$$

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.

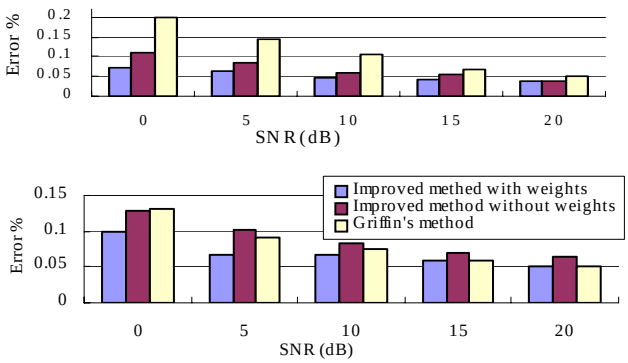

Figure 4-Comparison of errors of different pitch trackers for speech in train noise (top) and car noise (bottom) from 0dB SNR to clean.

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

$$Harmonicity = \frac{1}{NH \times N_{frames}} \sum_{N_{frames}} \sum_{k=1}^{NH} 10 \log \frac{P_k + P_{k+1}}{2P_{k,k+1}} \qquad (16)$$

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 MMSE system. It is evident that the proposed 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. Figure 8 presents two examples of the spectrograms of the original noisy and restored heritage archived speech records. These examples are more than a century old recordings of Florence Nightingale and Mozafaraldine Shah of Iran. The combined result of denoising the speech followed by bandwidth extension provides a more pleasant sounding output which is significantly better than other methods such as MMSE. Examples of contaminated and restored speech files are also available for hearing on http://dea.brunel.ac.uk/cmsp/florence_nighingale.htm.

8. CONCLUSION

This paper presented a parameter-tracking LP model combined with a harmonic and noise model of the excitation for restoration of the old archived 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 restoration system shows that it delivers improved results compared to MMSE method with significantly less artifacts such as musical noise.

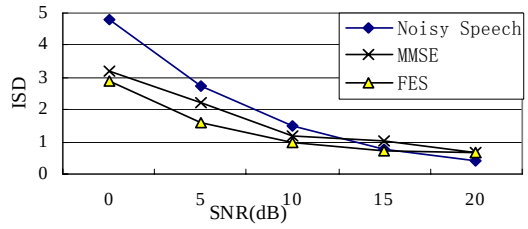

Figure 5 - Comparison of ISD of noisy speech in train noise pre-cleaned with MMSE and improved with formant-base enhancement system (FES) at SNR = 0, 5, 10, 15 dB.

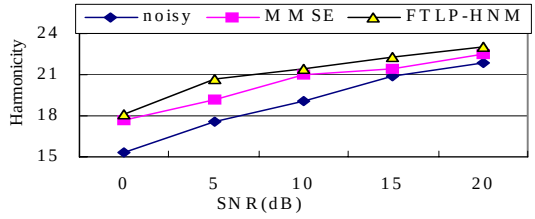

Figure 6-Comparison of harmonicity of MMSE and FTLP-HNM systems on train noisy speech at different SNRs.

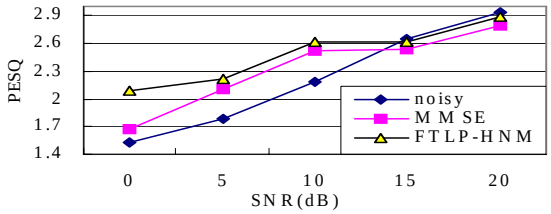

Figure 7-Performance of MMSE and FTLP-HNM on noisy train speech at different SNRs.

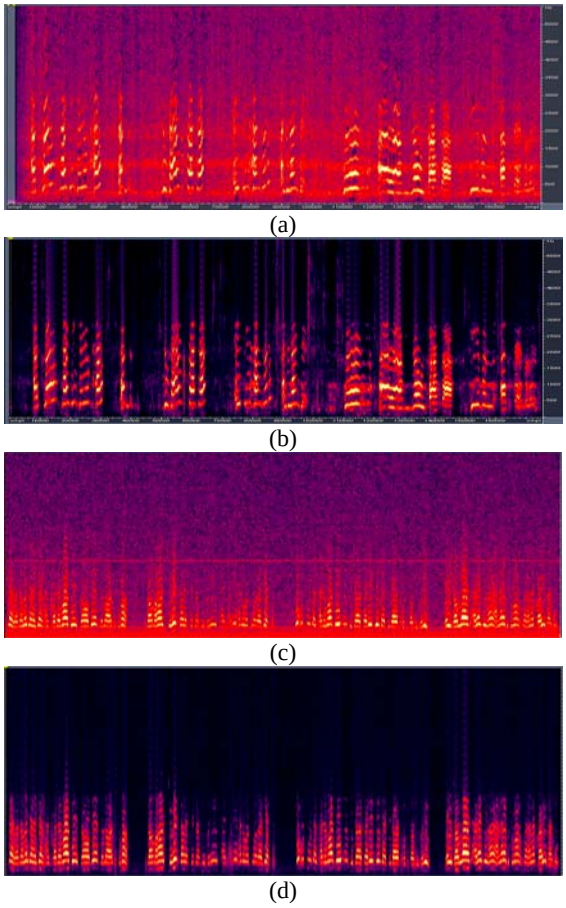

Figure 8- Two examples of noisy and restored archived speech (a,b) Florence Nightigale1-1890, (c,d) Mozafaraldine Shah of Iran-1906.

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. (2006).

[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).