Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing

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

a Institute of Acoustics, Chinese Academy of Science, Beijing, China
b Department of Electronic and Computer Engineering, Brunel University, UB8 3PH, UK
c School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK
d Institute of Sound and Vibration Research, University Road, Highfield, Southampton S017 1BJ, UK

Received 11 January 2006; received in revised form 23 November 2006; accepted 25 November 2006
Available online 26 December 2006

Abstract

This paper presents a formant tracking linear prediction (LP) model for speech processing in noise. The main focus of this work is on the utilization of the correlation of the energy contours of speech, along the formant tracks, for improved formant and LP model estimation in noise. The approach proposed in this paper provides a systematic framework for modelling and utilization of the inter-frame correlation of speech parameters across successive speech frames; the within frame correlations are modelled by the LP parameters. The formant tracking LP model estimation is composed of three stages: (1) a pre-cleaning spectral amplitude estimation stage where an initial estimate of the LP model of speech for each frame is obtained, (2) a formant classification and estimation stage using probability models of formants and Viterbi-decoders and (3) an inter-frame formant de-noising and smoothing stage where Kalman filters are used to model the formant trajectories and reduce the effect of residue noise on formants. The adverse effects of car and train noise on estimates of formant tracks and LP models are investigated. The evaluation results for the estimation of the formant tracking LP model demonstrate that the proposed combination of the initial noise reduction stage with formant tracking and Kalman smoothing stages, results in a significant reduction in errors and distortions.

© 2006 Elsevier Ltd. All rights reserved.

1. Introduction

De-noising speech enhances the quality and the intelligibility of voice communication in noisy environments. Noise reduction is particularly important for voice-based automated services and for users of mobile phones and hands-free phones in noisy moving environments such as in cars, trains, streets, conference halls and other potentially noisy venues. This paper describes a formant tracking linear prediction (LP) model for
Nomenclature

Symbols

\( y(m) \) noisy speech  
\( x(m) \) clean speech  
\( n(m) \) noise  
\( Y(l) \) DFT of the noisy speech  
\( X(l) \) DFT of the clean speech  
\( N(l) \) DFT of the noise  
\( X_l \) short-time spectral amplitude of clean speech  
\( \hat{X}_l \) an estimate of the short-time spectral amplitude of clean speech \( X_l \)  
\( G_l \) a spectral gain factor  
\( G(m) \) gain of the LP model  
\( r_k(m) \) time-varying radii  
\( \phi_k(m) \) angular frequencies of the poles of the LP model  
\( v_k \) formant feature vector  
\( F_k \) frequency of the formants  
\( B_k \) bandwidth of the formants  
\( M_k \) magnitude of the formants  
\( k \) index of the formants  
\( \Delta F_k(m) \) velocity of the \( k \)th formant at frame \( m \)  
\( O_n \) observations of the resonance frequencies of a speech frame  
\( F_k \) estimate of the \( k \)th formant  
\( A_q \) an HMM of the formants of phoneme \( q \)  
\( N \) number of formants  
\( \hat{A}_k \) a Gaussian mixture model of the \( k \)th formant  
\( F_{i,k}(m) \) frequency of the \( i \)th pole classified as the \( k \)th formant  
\( w_{k,i}(m) \) probability that the \( i \)th pole frequency is labeled as the \( k \)th formant  
\( \hat{F}_k(m) \) formant tracks of clean speech  
\( \hat{F}_k(m) \) formant tracks of noisy speech  
\( \text{SNR}_l \) signal to noise ratio at the discrete frequency \( l \)  
\( \text{SNR}_{\text{Thresh}} \) a signal noise ratio threshold  
\( I(\cdot) \) gamma function  
\( I_n(\cdot) \) Bessel function of order \( n \)  
\( F_{i,k}(m|m-1) \) a prediction of \( F_{i,k}(m) \) from estimates of the formant track up to and including time \( m-1 \)  
\( P(m) \) formant estimation error covariance matrix  
\( P(m|m-1) \) formant prediction error covariance matrix  
\( K(m) \) Kalman filter gain  
\( R \) measurement noise covariance matrix  
\( Q \) covariance matrix of the process noise

Abbreviations

LP linear prediction  
DFT discrete Fourier transform  
SS spectral subtraction  
MMSE minimum mean squared error  
MAP maximum a posteriori  
HMM hidden Markov model  
AR autoregressive  
LPSS LP spectral subtraction
improved speech processing in noise. Formant tracking LP models have applications in speech enhancement, speech coding, speech recognition and speaker identification.

An essential aspect of a speech de-noising method, based on an LP model, is the estimation of the spectral envelope, or equivalently the correlation matrix of clean speech, from which an LP model and the formants of speech can be estimated. The main motivation for this work is the modelling and utilization of the temporal trajectories of LP models along the formants for reducing the effect of noise, uncertainty and processing artifacts. Note that in this paper only additive noise is considered.

Formant tracking LP models provide a suitable framework for modelling the non-stationary temporal correlation of speech across successive frames. The choice of formant tracking LP models for enhancement of noisy speech follows from the common use of LP models for speech coding, formant tracking and speech enhancement (Schroeder, 1999; McCandless, 1974; Chen et al., 2000). The main issues in formant-based noisy speech processing are: (a) the choice of formant features; in this work the formant features, extracted from the LP model of speech, are the frequencies, bandwidth and magnitude of resonance at formants, (b) statistical models of formants such as hidden Markov models and Kalman filters, (c) de-noising and restoration of the formant tracks from noisy speech and (d) the use of de-noised formant features for applications such as speech enhancement.

Formants are the resonances of the vocal tract and their trajectories describe the contours of energy concentrations in time and frequency. At formants the signal to noise ratio is relatively higher than average and furthermore much of the discriminative information regarding phonemic labels and some of the speaker characteristics are encoded in the spectral features at formants (Lim and Oppenheim, 1978). Although formants are mainly defined for voiced speech, characteristic energy contours also exist for unvoiced speech as concentrations of energy at relatively higher frequencies.

Numerous signal processing techniques for formant track estimation have been developed over the past two decades. Popular alternative approaches include frequency domain techniques such as LP models that track the trajectories of the poles that model the resonance at formants (Deller et al., 1993) and peak-picking in short-time frequency spectrum (Shafer and Rabiner, 1970). Kopec’s original work is one of the first methods for formant tracking based on linear prediction models and hidden Markov models (HMMs) (Kopec, 1986). Kopec developed formant HMMs with discrete observation probabilities for modelling single or multiple formant tracks. The main difference between Kopec’s formant estimation and the work reported here are: (a) this work addresses the problem of formant estimation in noise whereas Kopec’s and most other formant estimation methods are developed for clean speech and (b) this work uses a 2D HMM structure with continuous probability densities.

In recent years, there has been extensive research aimed at improving the accuracy of formant estimation methods. A special session on format estimation in ICASSP2004 describes improvements to formant tracking including issues such as the use of linguistic classes and segmentation/labeling (Zheng and Hasegawa-Johnson, 2004) and the improved modelling of formant tracks for consonants (Lee et al., 2005). In Darch et al. (2005) the joint distributions of formants and cepstrum vectors are used for prediction of formants. For noisy speech, Chen and Loizou (2004) proposed a formant tracking method based on Wiener filtering of speech segments followed by a peak-picking algorithm.

Formant classification can be achieved through a rule-based classification method (Kalman, 1960; Rigoll, 1986; Mack and Jain, 1985) or via a probabilistic Bayesian framework using Gaussian mixture models (GMMs) or hidden Markov models (HMMs) of formants (Rentzos et al., 2003; Weber et al., 2001; Vergin
et al., 1996; Kim and Sung, 2001). The formant tracks obtained from formant classification of successive frames of speech can be modelled and smoothed using Kalman filters.

The focus of this paper is formant estimation in noise. Like most discrete-time speech processing systems, speech samples are segmented into overlapping frames with a frame duration of about 20–30 ms. Assuming that the noise is additive, the noisy speech, \( y(m) \), may be expressed in the time domain as the sum of the clean speech, \( x(m) \), and the noise, \( n(m) \), as

\[
y(m) = x(m) + n(m)
\]

where \( m \) is the discrete-time variable. In the frequency domain the discrete Fourier transform (DFT) of a segment of \( M \) samples of noisy speech is expressed as

\[
Y(l) = X(l) + N(l) \quad l = 0, \ldots M - 1
\]

where the complex variables \( Y(l) \), \( X(l) \) and \( N(l) \) are the DFTs of noisy speech, clean speech and noise, respectively, and \( l \) is the discrete-frequency variable. Eq. (2) can be expressed in the complex polar form in terms of the magnitude and phase of speech and noise as

\[
Y_l e^{j\theta_l} = X_l e^{j\theta_l} + N_l e^{j\theta_l} \quad l = 0, \ldots M - 1
\]

where \( Y_l \) and \( \theta_l \) are the magnitude and phase of the complex variable \( Y(l) = Y_l e^{j\theta_l} \).

Speech de-noising systems are often based on short-time spectral amplitude estimation methods which assume that the spectral samples are independent and identically distributed (IID). Examples are the spectral subtraction (SS) method (Vaseghi, 2005) and the Bayesian minimum mean squared error (MMSE) spectral amplitude estimation method (Ephraim et al., 1995). These estimators may be expressed by a simple equation as

\[
\hat{X}_l = G_l Y_l \quad l = 0, \ldots M - 1
\]

where \( \hat{X}_l \) is an estimate of the short-time spectral amplitude of clean speech \( X_l \) and \( G_l \) is a spectral gain factor that may be obtained from Wiener filter theory or from Bayesian estimation (Kim and Sung, 2001). In general the gain function, \( G_l \), is a function of the signal to noise ratio at frequency \( l \).

In its most basic form, the spectral amplitude estimator in Eq. (4) does not employ a model for the utilization of the correlation of spectral samples across frequency or time. The two main research issues in the processing of noisy speech are as follows:

(a) Modelling and utilization of the probability distributions and the intra-frame correlations of the speech and noise features within each noisy speech frame.

(b) Modelling and utilization of the probability distributions and the inter-frame correlations of speech and noise features across successive frames of noisy speech.

The main objective of this paper is to explore the use of formant tracking LP models for utilization of intra-frame and inter-frame correlation structures for improved speech model estimation in noise.

The classical methods for de-noising speech are mostly based on Wiener filter theory, which assumes that the signal and noise are stationary processes and requires the correlation matrices, or equivalently the power spectra, of the signal and noise (Vaseghi, 2005). Wiener filters employing LP models of speech and noise have been extensively employed for speech enhancement (Deller et al., 1993; Lim and Oppenheim, 1979). Examples of LP-based Wiener filters are the maximum a posteriori (MAP) probability speech enhancement method developed by Lim and Oppenheim (1979), MMSE-based spectral estimator for speech enhancement (Ephraim et al., 1995; Ephraim and Malah, 1985), iterative Wiener filters with inter-frame and intra-frame constraints by line spectral pair (LSP) transformation (Hansen and Clements, 1987) and LP-based spectral subtraction (Vaseghi, 2005). A simple and widely studied approximation to the Wiener filter is spectral subtraction (Vaseghi, 2005; Boll, 1979; McAulay and Malpass, 1980). In spectral subtraction the short-time spectral amplitude is estimated through subtraction of an estimate of the noise spectrum from that of noisy speech. The well-known problem with spectral subtraction is that it results in the appearance of annoying short bursts of noise (Lim and Oppenheim, 1978).

Speech is a non-stationary process as are many noise processes in mobile environments. A major issue in de-noising speech is the modelling and utilization of the time-varying distribution of the trajectories of speech...
and noise parameters across successive noisy speech frames. The inter-frame temporal variations of speech spectra are generally modelled using a Markovian process such as hidden Markov model (HMM) or Kalman filter. HMMs can be used to decode noisy speech to obtain the most likely estimates of the power spectra of speech and noise for the implementation of Wiener filters (Sameti et al., 1998; Ephraim, 1992; Ephraim et al., 1989). Whereas HMMs provide a finite-state model of the time-varying probability distributions of speech and noise, Kalman filters provide a Markovian model of the non-stationary trajectories of the speech and noise across successive noisy speech frames (Gibson and Koo, 1991). Kalman filters have been applied to de-noising speech, notably in Gannot et al. (1998) and Paliwal and Basu (1987).

The reminder of this paper is organized as follows. Section 2 provides an analysis of noise and its effects on the observed distributions of formants and on formant estimation. Section 3 presents an overview of the proposed method for the estimation of formant tracking LP models. Section 4 describes methods of pre-cleaning the spectral amplitude of speech prior to the formant estimation stage. Section 5 describes the formant classification and formant track estimation process. Section 6 describes the use of Kalman filters for smoothing the trajectory of formant tracks. Section 7 describes performance measures for speech enhancement and presents evaluation results. Finally, Section 8 concludes the paper.

2. Analysis of noise and its effect on formant estimation

This section presents an analysis of the spectral characteristics of background noise in car, train and street environments and investigates the detrimental effects of noise on formant estimation.

2.1. Noise models

This subsection provides an overview of the distributions of the means and standard deviations of the spectrum of different types of noise and describes the methods commonly used for modelling noise. The types of noise we consider here are a BMW 3-series car traveling at 112 km/h on a motorway, train noise and street noise in central London recorded by colleagues from our laboratory. Fig. 1 shows the variation of the magnitude LP-spectrum of noise, and the corresponding means and standard deviations. The LP-spectrum of noise is obtained from magnitude frequency responses of LP models of noise frames. Typically a noise recording would provide many frames (a frame here has a duration of 25 ms). Fig. 1 shows that, for the particular noise databases that we have used, the LP-spectrum of car noise exhibits a range of variations within a band of about 5 dB around the average. The train noise and street noise exhibit higher variations.

The sources of the variations of noise, and possible models, are as follows:

(a) The random nature of noise may be modelled with a covariance matrix or a probability model.
(b) The non-stationary characteristics of noise sources can be modelled using dynamic models such as a Kalman filter and/or finite-state probabilistic models such as GMMs or HMMs.
(c) The multiplicity of noise sources; for example in a street environment there may be many different sources of noise. The multiplicity of noise can be modelled by a set of GMMs or HMMs.
(d) The uncertainty in feature extraction methods, such as DFT and LP methods, due to such factors as the length and position of windows, the number of pitch periods in each window and the model order.

The simplest noise model, used in spectral subtraction, is a template of the average magnitude spectrum of noise obtained from speech-inactive periods. In general the performance of noise reduction systems improves with the increasing utilization of statistical models of speech and noise from spectral subtraction to Wiener filter, Kalman filter and ultimately to Bayesian models of speech and noise such as HMMs.

2.2. The effect of noise on the estimation of formants

The database used to investigate the effect of noise on formants is the Wall Street Journal (WSJ) speech database. The speech signal is degraded by either car noise or train noise with an average SNR in the range from 0 to 20 dB. Formant tracks of both clean and synthesized noisy speech are obtained via LP-based formant
extraction and HMM probability models presented in Section 5. 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

$$\text{FSNR}(k) = 10 \log_{10} \left[ \frac{\sum_{l \in k(F_k - B_k/2)} X_l^2}{\sum_{l \in k} N_l^2} \right]$$

where $X_l$ is the magnitude spectrum of clean speech, $N_l$ the magnitude spectrum of noise and $F_k$ and $B_k$ are the frequency and bandwidth of the $k$th 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 average SNR, which may be a contributing factor to the fact that humans can recognize speech under severe noisy conditions.

Fig. 3 illustrates the effect of train and car noise on the histograms of the resonant frequencies of the poles of an LP model for the vowel $iy$. The LP model order is set to 13. The histogram was obtained from all examples of the phoneme $iy$ in the WSJ database for one speaker. Each peak of the histogram is indicative of the
existence of a formant. It can be seen that the effect of train noise is to introduce a spectral peak due to noise in
the vicinity of the first formant and consequently the actual first formant cannot be seen clearly. The noise also
broadens the bandwidth of the observations of speech resonances at formants. In general the effects of random
noise on the observations of the spectral features at formants are as follows:

(a) a broadening of the observed bandwidths of the resonances at formants,
(b) a shift in the observed formants towards the concentrations of the noise energy in the vicinity,
(c) introduction of new spectral peaks due to the mechanical vibration/rotations of the noise source.

The effect of noise on the observations of formant frequencies of vowels at different SNRs are shown in Fig. 4,
which displays the average of formants of vowels, obtained from training formant HMMs on noisy speech. Note that the first formant is most affected by noise due to a greater concentration of the energy of the noise
at its vicinity. Fig. 5 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 an LP model spectrogram of clean speech. The formant tracks are
obtained from HMMs described in Section 5. As expected, due to the relatively broader spread of train noise
energy in the frequency domain compared to that of car noise, formant track estimates from noisy car speech
are less affected than those from train noise. Note that the first formant track, which is the closest to the con-
centration of noise in the frequency domain, is most affected by noise and this effect is more pronounced in
train noise than in car noise.

To quantify the effects of noise on speech, an average formant track error is defined as the normalized dif-
fERENCE between the formant tracks obtained from clean (reference) speech and noisy speech:

$$E_k = \frac{1}{L} \sum_{m=1}^{L} \left[ \frac{F_k(m) - \bar{F}_k(m)}{F_k(m)} \right] \times 100\% \quad k = 1, \ldots, N$$  (6)
where $F_k(m)$ and $\hat{F}_k(m)$ are the formant tracks of clean and noisy speech respectively, $m$ the frame index and $L$ is the number of frames over which the formant error is measured. We should mention here that our ‘ground truth’ experiments have shown that the reference formant estimates from clean speech correspond well to the estimates visually tracked from spectrograms. Fig. 6 shows the percentage formant track errors for speech

Fig. 4. Variations of the average estimates of the formant frequencies observed in train noise at SNRs from 0 to 20 dB. The average values of formants were obtained from HMMs trained on formants extracted from noisy speech. The figures allow comparisons of formant estimates of noisy speech with those from clean speech.

Fig. 5. 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 the LP spectrogram of clean speech. Solid lines are formant tracks of clean speech; dash lines: formant tracks in train noise; dot lines formant tracks in car noise.

Fig. 6. Average (%) estimation error of speech formant tracks in train noise at different SNRs using an HMM formant tracker.
degraded with train noise. The results are averaged over 135 speech sentences. It is evident from Fig. 6 that HMM formant trackers are not robust and that formant tracking performance degrades with decreasing SNR.

3. Formant tracking LP model in noise

As shown in Section 2, noise degrades, and in some cases obscures, formant tracks. In this section a more robust formant tracking method is introduced that combines pre-cleaning of the spectral amplitude of speech with formant estimation and subsequent smoothing of the formant trajectories with Kalman filters.

The effect of additive noise on the DFT spectrum of speech is additive. However, as is evident from the histogram of formants in noise, Fig. 3, the effect of additive noise on the estimate of the frequency and bandwidth of the resonance at formants is rather complicated and depends on the shape of the noise, as well as the LP model order and cannot be expressed with a simple additive relationship. This complicates the modification of HMMs of formants to include the effect of noise. Hence, in the formant tracking method proposed here the following three-stage de-noising process is employed:

At the first stage of the de-noising process, an initial estimate of the short-time spectral amplitude of speech is obtained. This would leave behind some random residual noise and distortion that would have a detrimental effect on the subsequent formant track estimation stage, producing formant tracks that are noisy. At the second stage of the de-noising process an LP analysis of the amplitude spectrum is used to obtain an estimate of the poles of speech spectrum (Makhoul, 1975) which are then used as the raw data for formant classification and tracking. The role of the classifier is to label each pole with a formant. This may be achieved using a probability model (e.g. HMMs) of formants when such a model is available (Rentzos et al., 2003) or it may involve the use of a rule-based formant classification method as described in Section 3. This is followed by the third and final stage of formant estimation; a Kalman filtering of the formant trajectories to mitigate the effects of the residual noise and distortions left behind by the initial amplitude estimation process.

The proposed three-stage formant tracking LP model is illustrated in Fig. 7. At the front end there is a noise detection/estimation module for the detection of speech inactive periods during which the noise model is estimated and/or updated. The output formant estimates are then used to obtain an improved estimate of the LP model parameters along the smoothed formant tracks.

4. Pre-cleaning of spectral amplitudes of noisy speech

Prior to formant estimation, a pre-cleaning of the spectral amplitude of noisy speech is performed. In this section two pre-cleaning methods based on linear prediction spectral subtraction (LPSS) (Vaseghi, 2005) and minimum mean squared error (MMSE) spectral amplitude estimation are proposed (Ephraim et al., 1995).

In spectral subtraction an estimate of the average spectral magnitude of noise is subtracted from the noisy speech spectrum. For spectral subtraction (Schroeder, 1999; Vaseghi, 2005), the gain function of Eq. (4) is given by

\[
G_{SS}(l) = \begin{cases} 
1 - \alpha(f)\hat{N}(l)/Y(l) & \text{if } \text{SNR}_{\text{Thresh}} > \text{SNR}_l \\
\gamma \exp(-\beta(\text{SNR}_{\text{Thresh}} - \text{SNR}_l)) & \text{else}
\end{cases}
\]  

(7)

Q. Yan et al. / Computer Speech and Language 21 (2007) 543–561
where SNR<sub>j</sub> is the signal to noise ratio at the discrete frequency <i>j</i>, SNR<sub>Thresh</sub> a signal noise ratio threshold, α(<i>j</i>) a subtraction factor and γ and β are control variables that can be used to ensure continuity at the switching point in Eq. (7). In this work a version of the spectral subtraction method based on the LP model of noisy speech is used (Vaseghi, 2005), where <i>Y</i>(<i>j</i>) and <i>N</i>(<i>j</i>) are the frequency responses of LP models of noisy speech and noise, respectively (Ephraim et al., 1995).

The MMSE method of estimation of the spectral amplitude of speech (Vaseghi, 2005) is a Bayesian estimation method employing a mean squared error cost function. In the MMSE formulation of Ephraim and Malah (1985), it is assumed that the prior probability density function of the complex spectrum of clean speech is Gaussian. This assumption leads to a Rayleigh probability density function for the magnitude spectrum of clean speech and a uniform probability density function for its phase. It is further assumed that the complex spectrum of noise has a Gaussian probability density function. The MMSE gain function is given by

\[ G_{\text{MMSE}}(l) = I(1.5) \frac{\sqrt{\beta_i}}{\gamma_i} \exp \left( -\frac{\beta_i}{2} \right) \left[ (1 + \beta_i)I_0 \left( -\frac{\beta_i}{2} \right) + \beta_iI_1 \left( -\frac{\beta_i}{2} \right) \right] \]

where \( I(\cdot) \) is the gamma function, \( I_d(\cdot) \) is the Bessel function of order <i>d</i> and the parameters \( \beta_i \) and \( \gamma_i \) are defined as

\[ \beta_i = \frac{\xi_i}{\gamma_i}, \quad \gamma_i = \frac{\sigma_X^2(l)}{\sigma_N^2(l)}, \quad \beta_i = \frac{Y^2(l)}{\sigma_N^2(l)} \]

where \( \sigma_X^2(l) \) and \( \sigma_N^2(l) \) are the variance of speech and noise spectra, \( \xi_i \) is known as the prior signal to noise ratio and \( \gamma_i \) is known as the posterior signal to noise ratio.

Fig. 8 provides a comparison of the performance of linear prediction spectral subtraction (LPSS) and MMSE methods of spectral amplitude estimation. The performance measures used here are the Itakura-Saito distance (ISD) and signal to noise ratio (SNR) whose calculations are explained in Section 7.1. It is evident that the MMSE consistently performs better than the spectral subtraction method for most SNRs.

5. Formant estimation

This section presents the method for the estimation of formant parameters. An LP analysis of speech is used to obtain a first estimate of the formant feature candidates. From the formant feature candidates a two-dimensional HMM (Rentzos et al., 2003; Weber et al., 2001; Vergin et al., 1996; Kim and Sung, 2001) is trained and used to model and track the variation of formants in both time and frequency. In this work the formants are obtained from the poles of a linear prediction model of speech. In the \( z \)-transform domain, a linear prediction model of speech \( X(z) \) may be expressed as

\[ X(z) = E(z)V(z) \]

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

\[ V(z, m) = G(m) \frac{1}{1 + r_0(m)z^{-1}} \prod_{k=1}^{n/2} \frac{1}{1 - 2r_k(m) \cos(\varphi_k(m))z^{-1} + r_k^2(m)z^{-2}} \]
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 for frame $m$. In Eq. (11) speech is modelled by a cascade of time-varying second order resonator models of the formants and also 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 the speech.

The poles obtained from the LP model are considered as formant candidates and are subsequently represented as a formant feature vector, $v_k$, comprising the frequency, $F_k$, bandwidth, $B_k$ and magnitude, $M_k$, of the formants together with their velocity derivatives,

$$v_k = \left[ F_k, B_k, M_k, \Delta F_k, \Delta B_k, \Delta M_k \right] \quad k = 1, \ldots, N \tag{12}$$

where the number of formants is typically set to $N = 5$ at a sampling rate of 10 kHz. Velocity derivatives are denoted by $\Delta$ and are computed as the slope of the features over time. For example $\Delta F_k(t)$, the velocity of the $k$th formant at frame $t$, is obtained as

$$\Delta F_k(t) = \frac{\sum_{m=1}^{L} m(F_k(t + m) - F_k(t - m))}{\sum_{m=1}^{2} m^2} \quad k = 1, \ldots, N \tag{13}$$

where $L$ is the number of frames over which the slope of the trajectory of formants is estimated. There are three issues in the accurate modelling of formants; (i) modelling the distribution of formants using a probability model such as an HMM or Gaussian mixture model (GMM), (ii) estimating formant tracks with a Viterbi decoder and (iii) smoothing the trajectory of formants with Kalman filters. This section considers the application of HMMs for modelling the probability distributions of formants and for estimating formant tracks. The smoothing of formant tracks with Kalman filters is considered in Section 6.

### 5.1. Two-dimensional HMM-based formant tracking

The probability distributions of the trajectories of formants across frequency and time can be modelled by two-dimensional, frequency-time, HMMs (2D-HMM) described in detail in Rentzos et al. (2003), Weber et al. (2001), Vergin et al. (1996) and Kim and Sung (2001). An overview of 2D-HMM of formants is given in this section. Fig. 9 illustrates a 2D-HMM of formants of a phonemic unit where as it is shown, each state may
include a Kalman filter for tracking the trajectory of the formant within the state. A 2D-HMM is a combination of an HMM along time and an HMM along frequency. The HMM along the time dimension (temporal HMM) divides speech frames into sub-phonetic segments associated with the HMM states. The states of the HMM along the frequency domain (formant HMM) model the distribution of the formants of speech associated with each sub-phonetic state of the temporal HMM.

The purpose of using an HMM along frequency is to allow the observations of each formant to be associated with the distribution of a distinct formant state. For example the observations of the \( k \)th formant are associated with the \( k \)th state of the HMM along the frequency domain. The use of a 2D-HMM structure allows the application of established HMM methods and software for modelling the distribution of formants. Note that, in general, the formants coexist simultaneously in different states of formant HMMs and that there is in fact no actual transition from one formant state to the next state. However, the use of a left–right HMM structure for formants is useful for constrained classification of the poles of LP model (sorted in order of increasing frequency of resonance of poles) into the formants for those cases when the number of LP model poles exceeds the number of formants.

For training 2D-HMMs, first, conventional (temporal) HMMs, trained on a database of cepstral speech feature vectors, are used to model phonemes and estimate the sub-phonetic segment boundaries of training speech associated with each state of the HMMs. Then the formant feature vectors from the speech segments associated with each sub-phonetic state of the temporal HMMs are used to train formant HMMs. The training of 2D HMMs is described in detail in Rentzos et al. (2003).

Along the frequency axis, a GMM in each state of a 2D-HMM models the probability distribution of the formants associated with that state. The 2D-HMMs are trained with formant feature vectors obtained from Eqs. (12) and (13). Given a set of observations of the resonance frequencies of a speech frame \( O_n \), the maximum likelihood (ML) decoder of the associated formants is obtained as

\[
[F_1, F_2, \ldots, F_N] = \arg \max_{F_1, F_2, \ldots, F_N} P(O_n, [F_1, F_2, \ldots, F_N]|A_q) \quad k = 1, \ldots, N
\]

where \( O_n \) is obtained from the poles of an LP analysis of a frame of a speech phoneme and sorted in terms of the increasing frequency, \( F_k \) the ML estimate of the \( k \)th formant, \( A_q \) is an HMM of the formants of phoneme \( q \) and \( N = 4–6 \) is the number of formants.

Eq. (14) is implemented with a Viterbi decoder which provides a maximum likelihood estimate of the discrete formant state sequence given the observation vector and the HMM models. The Viterbi method labels the observations with discrete formant label identifiers (1 to \( N \)) and calculates the actual formant values. Hence, the estimate of the \( k \)th formant is obtained from the ML state using the probability-weighted average of the mean values of the models of the Gaussian mixture distributions associated with the ML state.

HMMs are used to classify the poles of the LP model into formants and estimate formant trajectories. The HMM-based formant classifier may associate two or more formant candidates (poles of LP model) \( F_{k1}, F_{k2} \) with the same formant label \( k \). In these cases formant estimation is achieved through minimization of a weighted MMSE objective function as Ho et al. (2002)

\[
\hat{F}_k(m) = \arg \min_{F_k(m)} \sum_{i=1}^{N_k(m)} w_{ki}(m) \left( \frac{(F_{ki}(m) - F_k(m))^2}{B_{ki}(m)^2} \right) \quad k = 1, \ldots, N
\]

where \( F_{ki}(m) \) is the frequency of the \( i \)th pole classified as the \( k \)th formant, \( w_{ki}(m) = P(F_{ki}(m)|\lambda_k) \) is the probability that the \( i \)th pole frequency is labeled as the \( k \)th formant, \( \lambda_k \) is a Gaussian mixture model of the \( k \)th formant and \( N_k(m) \) is the total number of poles of the \( k \)th speech frame classified as formant \( k \). In Eq. (15) the distance of each pole frequency candidate from the formant estimate is weighted by a probabilistic weight \( w_{ki}(m) \) and a perceptual weight \( 1/B_{ki}^2 \) where \( B_i \) is the formant bandwidth. Note that \( N_k(m) \) is usually one or two. For the case when \( N_k(m) = 1 \) then \( \hat{F}_k(m) = F_{k1}(m) \). For the case when \( N_k(m) = 2 \), taking the derivative of Eq. (15) with respect to \( F_{k1}(m) \) yields an MMSE interpolated estimate of the \( k \)th formant at time \( m \) as

\[
\dot{F}_k(m) = \frac{\alpha_{k1}}{\alpha_{k1} + \alpha_{k2}} F_{k1}(m) + \frac{\alpha_{k2}}{\alpha_{k1} + \alpha_{k2}} F_{k2}(m) \quad k = 1, \ldots, N
\]

where \( \alpha_{ki} = w_{ki}(m)/B_{ki}^2(m) \).
5.2. Rule-based formant tracking

For speech processing applications, where probability models of formants (such as HMM of formants) are not available, rule-based methods of formant classification and tracking may be used as an alternative. The rules for classification of each pole with a formant index are based on the Euclidean distance of the pole from the most recent estimate of the trajectories of the formants. A simple nearest neighbor classifier for estimation of the index \( k \) of the formant associated with \( i \)th pole observation \( O_i(m) \) may be expressed as

\[
k = \arg \min_j (|O_i(m) - F_j(m - 1)|) \quad j = 1, \ldots, N
\]

where the classification of an observation \( O_i(m) \) is based on its distance from the estimate of formant values at time \( m - 1, F_j(m - 1) \). If two poles are associated with the same formant then a weighted mean of the poles is obtained using Eq. (16) where the weight for each pole may be obtained from

\[
w_{ki}(t) = \exp(-\gamma_i O_i - F_i) \gamma_i^2,
\]

where \( \gamma_i \) can be set to the inverse of an estimate of the variance of \( F_i \) to produce a measure that is similar to a Gaussian probability.

6. Formant track smoothing with state-dependent Kalman filters

Kalman filters are Markov processes that can be used for the modelling, prediction and smoothing of the trajectory of a random process such as that of a formant track (Kalman, 1960). The application of Kalman filters here requires a state-space model of the formant trajectory described in Section 6.1.

It is worth noting that after pre-cleaning speech with spectral amplitude estimation, the residual noise is due to the random variation around the spectral mean of the noise. In this work it is assumed that the residue noise manifests itself as additive disturbances on the formant tracks. Kalman filters are used to reduce the effect of residual noise on the estimates of the trajectory of formants.

6.1. Kalman filter state-space equations for formant trajectories

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

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

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 \( k \)th formant observation is obtained from the \( k \)th pole, \( p_k(m) \), modelled as

\[
p_k(m) = \hat{F}_k(m) + d_k(m)
\]

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

\[
\hat{F}_k(m|m-1) = \mathbf{C}\hat{F}_k(m-1) = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_p & c_{p-1} & c_{p-2} & \cdots & c_1
\end{bmatrix} \begin{bmatrix}
F_k(m-P) \\
F_k(m-P+1) \\
F_k(m-P+2) \\
\vdots \\
F_k(m-1)
\end{bmatrix}
\]

\[P(m|m-1) = P(m-1) + Q\]  

Measurement update equations

\[
K(m) = P(m|m-1)(P(m|m-1) + R)^{-1}
\]

\[
\hat{F}_k(m) = \hat{F}_k(m|m-1) + K(m)(p_k(m) - \hat{F}_k(m|m-1))
\]

\[
P(m) = (I - K(m))P(m|m-1)
\]

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

6.2. 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. As illustrated in Fig. 9, HMMs of speech can be integrated with Kalman filters to provide a finite-state Markov chain where in each state of the HMM the signal trajectory is modelled by a Kalman filter which itself is a Markovian model. State-dependent Kalman filters deal well with changes from voiced to unvoiced speech. Furthermore, the Kalman filter outputs can be interpolated over the transition regions to provide a more smoothed transition.

For cases where Kalman filter trackers are used with rule-based formant classification, instead of HMM-based classification, 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. It is worth noting that within voiced segments the formant trajectories are continuous and furthermore there is often some continuity between the formant trajectories of voiced speech on the two sides of an unvoiced segment.

7. Performance evaluation

The speech examples used for the evaluation of the performance of formant trackers and LP model estimation are a subset of five male speakers and five female speakers from the Wall Street Journal (WSJ) database. 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.

7.1. Speech SNR and distortion measurements

The distortion measures used here for the evaluations of the improvements in speech model estimates, are the segmental signal to noise ratio (SNR) and the Itakura-Saito distance (ISD) measure. The average SNR is defined as
SNR = 10\log_{10}\left(\frac{P_{\text{Signal}}}{P_{\text{Noise}}}\right) \text{ dB} \tag{25}

where $P_{\text{Signal}}$ and $P_{\text{Noise}}$ are the power of signal and noise respectively. The segmental SNR is defined as

$$
\text{SNR}_{\text{seg}} = \frac{1}{L} \sum_{m=1}^{L} 10 \times \log_{10} \left( \frac{\sum_{l=1}^{N} X_m^2(l)}{\sum_{l=1}^{N} (X_m(l) - \hat{X}_m(l))^2} \right) \text{ dB} \tag{26}
$$

where $X_m(l)$ and $\hat{X}_m(l)$ are the clean signal and restored signal at frame $m$, $L$ the total number of frames and $N$ is the number of samples in each frame. Fig. 10 shows the variation of segmental SNR at average SNRs of 0, 5 and 10 dB in car noise. It can be seen that the segmental SNR of speech signals can fluctuate widely about the average value.

The Itakura-Saito Distance measure (Deller et al., 1993; Turunen and Vlaj, 2001) is defined as

$$
\text{ISD}_{12} = \frac{1}{L} \sum_{m=1}^{L} (a_1(m) - a_2(m))^T R_1(m) (a_1(m) - a_2(m))
$$

where $a_1(m)$ and $a_2(m)$ are the linear predication model coefficient vectors calculated from clean and transformed speech at frame $j$ and $R_1(m)$ is an autocorrelation matrix derived from the clean speech. Due to the asymmetry of ISD measure (i.e. $\text{ISD}_{21} \neq \text{ISD}_{12}$) the following symmetric ISD measure is used

$$
\text{ISD}_{\text{sym}} = \frac{\text{ISD}_{12} + \text{ISD}_{21}}{2} \tag{28}
$$

The ISD criterion Eq. (27) 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 measure of Eq. (25).

### 7.2. Performance evaluation of formant tracking LP model

A formant-track percentage error measure, Eq. (6), is used for the evaluation of the performance of the Kalman-based formant tracker described in Section 6 for the restoration of the formants of noisy speech. The reference formant tracks are obtained via a 2D-HMM from the clean speech. The results of formant estimation with and without noise reduction and Kalman smoothing are shown in Fig. 11. The results show that the application of LPSS or MMSE results in a significant reduction of formant tracking error. Further improvement in formant track estimation is obtained through application of Kalman filters. Through the proposed noise reduction method, over 60% reduction in formant track error has been achieved in tracking the first formant, which is most affected by noise. In less affected higher formants (F2–F4), the Kalman-based method recovers the formant track with an average improvement of about 15%. To further verify the formant estimation results, Fig. 12 shows the percentage formant error for different methods with reference to a small

![Fig. 10](image-url) The distributions of SNRs of speech active and speech inactive frames at an average SNR of 0 dB, 5 dB and 10 dB in a car noise environment.
Fig. 11. Average % error of formant tracks of speech degraded with train noise and cleaned speech using LPSS, MMSE and Kalman filters, the results were averaged over five males and females.

Fig. 12. Average % error of formant tracks of speech degraded with train noise and cleaned speech using LPSS, MMSE and Kalman filters. These results were obtained with reference to hand corrected formants of clean speech. Ten sentences from one male and one female speaker were used.

Fig. 13. Comparison of clean formant tracks (red thin solid line) and cleaned formant tracks (blue dash dot) and noisy formant tracks (black thick solid line) for a segment of speech ‘time in years’. The background is a spectrogram of the LP model of clean speech. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 14. A block diagram illustration of the proposed speech enhancement system.
number of hand corrected formant tracks of clean speech. Note that the results shown in Fig. 12 agree with those in Fig. 11. Fig. 13 illustrates a comparative example of the improvements in formant tracking that result from the proposed noise reduction method where the formant tracks of clean and noisy speech and the restored formant tracks are superimposed on the spectrogram of clean speech.

Fig. 15. Comparison of SNR and ISD of speech degraded with train noise and pre-cleaned with linear prediction spectral subtraction (LPSS) and subsequently improved with formant-based enhancement system (FES) at SNR = 0, 5, 10, 15 dB. Silences are excluded here.

Fig. 16. Comparison of spectrograms of LP model from (a) clean speech ; (b) speech in train noise; (c) LPSS-cleaned speech; (d) MMSE-cleaned speech (e) formant-tracking enhancement with formants tracks obtained from LPSS pre-cleaning (f) formant-tracking enhancement with formants tracks obtained from MMSE pre-cleaning.
In this subsection, the formant tracking LP model is applied to the restoration of speech in noise using the de-noising system illustrated in Fig. 14. The speech signal is sampled at a rate of 10 kHz. Speech is divided into frames 25 ms long with a frame overlap of 10 ms. For LP analysis of the input signal, a model order of 13 is used. A voice activity detector (VAD) is employed to detect speech active/inactive periods. The noise model is estimated and updated from the speech-inactive, noise-only, periods. An amplitude estimation method, such as MMSE or LPSS is then used to pre-clean the noisy signal. This generally leaves behind some processing distortions in the form of short-duration bursts of noise known as ‘musical noise’ (Araki et al., 2005). Successive frames of pre-cleaned speech are then input to the formant tracking module, which incorporates formant classification followed by Kalman filters for filtering out residual noise and distortion as described in Section 4. The improved formant tracking LP model coefficients are then used for speech enhancement.

Fig. 15 shows the improvements in terms of both ISD and SNR distortion measures resulting from the combination of LPSS/MMSE with the formant tracking system. It is evident that the proposed formant-based speech enhancement system (FES) achieves a better (lower) ISD but a similar SNR compared to LPSS/MMSE alone in noisy conditions (below 20 dB). However, ISD is generally considered to be a better measure of speech quality than SNR (Deller et al., 2000).

Fig. 16 illustrates examples of spectrograms of noisy speech, speech de-noised by LPSS, speech de-noised by formant tracking enhancement system and clean speech. It is clear that the LPSS results in a number of fragments of spurious short bursts of tones in the enhanced speech (Fig. 16b) while in Fig. 16c the speech cleaned by the formant tracking method does not exhibit these distortions. The enhanced restoration of smoother trajectories of formants, and the rejection of short bursts of noise, is an important desirable aspect of the proposed formant tracking LP model estimation in noise.

8. Conclusion

This paper presented a formant tracking LP model for processing of noisy speech. The proposed formant tracking LP model provides a means for modelling the intra-frame and inter-frame correlations of speech frames. The model introduced in this paper provides a framework for the integration of LP models, probability models of formants (such as HMMs) and Kalman filters to obtain improved performance in noisy conditions. A formant track estimation method based on a combination of spectral amplitude estimation, formant classification and Kalman filtering achieves up to a 60% reduction in formant track estimation error. Evaluations of the de-noising system using two objective measures, namely ISD and SNR, show that the system achieves a similar SNR to the spectral amplitude estimation system but achieves a significantly better score in terms of ISD. However, we should mention here, as pointed out by one of the referees of this paper, that the true gain may be different in real conditions due to the influence of the Lombard effect (Hansen, 1994).

Further work on this noise processing system currently undertaken by the authors involves the integration of the formant tracking LP model with tracking of harmonics and noise components of the speech excitation of the LP model.

Acknowledgement

We wish to thank the UK’s EPSRC for funding Project No. GR/S30238/01.

References


