A FORMANT TRACKING LP MODEL FOR SPEECH PROCESSING IN CAR/TRAIN NOISE

Qin Yan  Esfandiar Zavarehei  Saeed Vaseghi  Dimitrios Rentzos
Department of Electronic and Computer Engineering
Brunel University, UK UB8 3PH
{Qin.Yan, Esfandiar.Zavarehei, Saeed.Vaseghi, Dimitrios.Rentzos}@brunel.ac.uk

ABSTRACT
This paper investigates the modeling and estimation of spectral parameters at formants of noisy speech in the presence of car and train noise. Formant estimation using two-dimensional hidden Markov models (2D-HMM) is reviewed and employed to study the influence of noise on observations of formants. The first set of experimental results presented show the influence of car and train noise on the distribution and the estimates of the formant trajectories. Due to the shapes of the spectra of speech and car/train noise, the 1st formant is most affected by noise and the last formant is least affected. The effects of inclusion of formant features in speech recognition at different SNRs are presented. It is shown that formant features provide better performance at low SNRs compared to MFCC features. Finally, for robust estimation of noisy speech, a formant tracking method based on combination of LP-spectral subtraction and Kalman filter is presented. Average formant tracking errors at different SNRs are computed and the results show that after noise reduction the formant tracking errors of 1st formant are reduced by 60%. The de-noised formant tracking LP models can be used for recognition and/or enhancement of noisy speech.

1. INTRODUCTION
Formants are resonance frequencies of speech. Most of the discriminative information conveying phonemic labels and also some of the speaker characteristics are encoded in spectral features at formants. In the context of noisy speech processing the use of speech features at formants is particularly interesting because the representation of speech by formant features may be viewed as a non-uniform data-adaptive sampling process where the feature samples at formants have relatively high signal to noise ratio and discriminative information content. The recovery of formant tracks from noisy speech can make a significant contribution to speech enhancement or speech recognition in noisy condition.

The main issues for formant-based noisy speech processing are: (a) the choice of formant features and models, (b) method of estimation of spectral features at formant tracks, (c) the use of formant features for speech recognition and enhancement and (d) evaluation of formant features and models compared to conventional features and models such as cepstrum features for speech recognition or spectral-subtraction for enhancement of noisy speech.

Numerous signal processing techniques for formant estimation have been developed. Popular approaches can be classified as frequency domain techniques such as peak-picking in short-time frequency spectrum [1] and parametric techniques such as Linear Prediction (LP) models [2]. Formant estimation becomes complicated in the presence of correlated background noise such as car and train noise as the spectrum of noise from revolving mechanical sources have their own spectral peaks that affect the number and positions of the observed peaks in noisy speech spectrum.

This paper concentrates on formant estimation for speech in noisy car/train environments. In section 2 the extraction of formant features and the probabilistic modeling of variation of formant features are reviewed. Section 3 investigates the effect of noise on formant models. In section 4 a method, for robust estimation of formants of noisy speech, based on a combination of LP-spectral subtraction and Kalman filter is described. Section 5 concludes the paper.

2. FORMANT ESTIMATION IN CLEAN SPEECH

2.1 Formant Track LP Models
Speech is modeled as

\[ X(z) = G(z)V(z)L(z) \]

where \( G(z) \), \( V(z) \) and \( L(z) \) are the z-transform of glottal pulse, vocal tract and lip radiation[3]. The vocal tract is modeled by a cascade of second order formant-tracking LP models as

\[ V(z, m) = G(m)\prod_{k=1}^{N} \frac{1}{1-2r_k \varphi_k (m) z^{-1} + r_k^2 (m) z^{-2}} \]

where \( \varphi_0(m) \), \( r_0(m) \) and \( G(m) \) are the time-varying frequency and radius of the poles and gain of a second order LP model. The spectral resonance at formant is characterized by a parameter vector comprising of the frequency \( F_k \), bandwidth \( BW_k \) and magnitude of the resonance \( I_k \) and their temporal slopes of variation as

\[ F_k = [F_k, BW_k, I_k, \Delta F_k, \Delta BW_k, \Delta I_k] \]

where \( \Delta \) denotes the slope of trajectory over time.

In this paper the glottal pulse and lip radiation effects are not explicitly modeled, although the effect of glottal pulse is mitigated through the use of a pre-emphasis filter.

2.2 2D-HMM Based Formant Track Estimation
The variations of formants’ trajectories of each phoneme are modeled by 2D-HMMs [4]. A 2D-HMM is a combination of a 1-D HMM along the time and a 1-D HMM along the frequency. Along the frequency axis, a Gaussian mixture model in each state of a 2D-HMM models the distribution of
one formant of the phoneme. 2D-HMMs are trained with formant feature vectors (Equation 3).

Given a set of observations of resonance frequencies \( O_n \), the maximum likelihood estimate of the associated formants is obtained as

\[
[\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_j] = \arg \max_{F_1, F_2, \ldots, F_j} P(O_n | [F_1, F_2, \ldots, F_j] | A_n)
\] (4)

Where \( O_n \) is obtained from the poles of an LP model of a segment of a speech phoneme and sorted in terms of increasing frequency. \( A_n \) is an HMM of the formants of phoneme \( m \) and \( N = 4 \) to 6 is the number of formants.

HMMs are used to classify formant candidates and estimate the trajectory of formants. The HMM-based formant classifier may associate two or more formant candidates \( F_{k,\phi} \) with the same formant \( b \). In these cases formant estimation is achieved through minimization of a weighted mean square error objective function\(^2\)

\[
\hat{F}_k(t) = \min_{I_k(t)} \sum_{i=1}^{I_k(t)} w_{b_k}(t) \left( \frac{(F_{k}(t) - F_{k}(t))^2}{B_{k}(t)^2} \right)
\] (5)

where \( t \) denotes the frame index, \( b \) is the formant index, \( I_{k,\phi} \) is the total number of the formant candidates classified as formant \( b \). The squared error function is weighted by a perceptual weight \( 1/(BW_i)^2 \) where \( BW_i \) is the formant bandwidth and a probabilistic weight defined as \( w_{b_k}(t) = P(F_k | A_b) \) where \( A_b \) is the Gaussian mixtures models of the \( b^\text{th} \) formant state of a phoneme-dependent HMM of formants.

3. EFFECT OF NOISE ON FORMANT MODEL

The databases used in the following analysis include 135 clean speech sentences from WSJ together, BMW car noise and train noise. The speech signals are down sampled from 16 kHz to 10kHz and degraded by either car or train noise with a global SNR in the range from 0 to 20dB. The noise have been recorded by colleagues from our lab. Formant tracks of both clean and synthesized noisy speeches are obtained via 2D HMM-based LP models reviewed in section 2.

3.1 The Effect of Noise on Formant Distribution

To quantify the effects of the noise on formants, formant signal to noise ratio (FSNR) is defined as

\[
FSNR(k) = 10 \log \left( \frac{\sum f \in F_k | X(f)|^2}{\sum f \in F_k | N(f)|^2} \right)
\] (6)

Where \( F_k \) and \( B_k \) are frequency and bandwidth of the formant \( k \) \( X(f) \) is the squared magnitude spectrum of clean speech while \( N(f) \) is the noise spectrum. Table 1 displays FSNR of car and train noisy speech. It is evident that SNRs at formants are much higher than the overall SNR, which may be one of the contributing reasons that we can recognize speech under severe noisy conditions since speech at resonance frequencies are relatively less affected by the noise.

A series of speaker-dependent speech recognition experiments are performed to compare speech recognition using formants and traditional MFCC in noisy conditions. Table 2 displays recognition rates using three-state

<table>
<thead>
<tr>
<th>SNR</th>
<th>MFCC</th>
<th>FORMANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>35.9</td>
<td>38.6</td>
</tr>
<tr>
<td>CLEAN SPEECH</td>
<td>88.5</td>
<td>77.7</td>
</tr>
</tbody>
</table>

Table 2: Phone recognition rate (%) of an American speaker at SNR = 0 and clean condition.

Figure 1: Effect of train noise on the vowel ‘iy’

Figure 2: average formant frequencies of an American speaker in train noise condition.
monophone speaker-dependent HMMs with 20 Gaussian mixtures per state. The formant features include the first five formant frequencies, their bandwidths, magnitudes and their temporal velocity (delta) and acceleration (delta-delta) features. The results indicate that formant features provide improved performance compared to MFCC in noisy conditions while the reverse holds in clean condition. This seems to agree with the observation that formants of speech are relatively less affected by noise.

The effects of noise on the spectral features at formants are as follows: (a) to broaden the apparent bandwidth of the existing formants, (b) to shift the apparent formant frequency due to noise energy concentration at the vicinity of formants and (c) to introduce new spectral peaks due to the vibration modes of the noise source.

Figure 1 illustrates the effect of train noise on formant distribution of vowel /iy/. It is shown that the effect of train noise is the appearance of a spectral peak due to noise in the vicinity of the first formant. The noise also broadens the bandwidth of observations of speech resonances. The effects of noise on formant observations can be seen clearer in Figure 2, where average formants of vowels are shown at difference SNR scenarios.

3.2 The Effect of Noise on Estimates of Formant Tracks

Figure 3 illustrates examples of formant tracks of speech in train and car noise at SNR=0. The formant tracks are obtained from 2D-HMM. As expected, due to the relatively broader spread of the energy of train noise in frequency domain compared to car noise, the estimates of formant tracks of noisy speech in car noise are less affected by noise than those from train noise, particularly in the first formant track. Further

<table>
<thead>
<tr>
<th>SNR</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>51.3</td>
<td>12.5</td>
<td>6.3</td>
<td>3.7</td>
<td>2.6</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>9.7</td>
<td>4.6</td>
<td>2.7</td>
<td>1.8</td>
</tr>
<tr>
<td>10</td>
<td>32.3</td>
<td>7.4</td>
<td>3.4</td>
<td>2</td>
<td>1.4</td>
</tr>
<tr>
<td>15</td>
<td>23.1</td>
<td>5.8</td>
<td>2.6</td>
<td>1.5</td>
<td>1.1</td>
</tr>
<tr>
<td>20</td>
<td>15.6</td>
<td>4.6</td>
<td>2.1</td>
<td>1.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Average errors (%) of formant tracks in train noisy speech

To quantify the affects of train/car noise on speech, an average formant track error measure, defined as the difference between formant tracks obtained for clean (reference) and noisy speech is calculated as follows.

$$E_k = \frac{1}{N} \sum_{m=1}^{N} \left[ \frac{F_k(m) - \hat{F}_k(m)}{F_k(m)} \right]$$

(7)

Where $F_k(m)$ and $\hat{F}_k(m)$ are the formant tracks of clean and noisy speech. In table 3 percentage formant track errors are averaged over 135 sentences contaminated with train noise. It is evident that 2D-HMM formant trackers are very sensitive to noise and that formant tracking performance degrades with the decreasing SNR.

4. FORMANT ESTIMATION IN NOISE

The limitation of format track estimation method based on 2D-HMM and LP models under noisy condition has been analyzed in previous section. In this section a more robust formant tracking method based on combination of LP spectral subtraction and Kalman filter are purposed.

4.1 Formant Tracking Using LP-based Spectral Subtraction and Kalman Filter

In LP-based spectral subtraction the LP spectrum of speech [5] is estimated as

$$\hat{X}_{LP}(f) = Y_{LP}(f) - \alpha(f) \hat{N}_{LP}(f)$$

(8)

where $Y_{LP}(f)$, $\hat{X}_{LP}(f)$ and $\hat{N}_{LP}(f)$ are the spectra of noisy signal, signal estimate and noise estimate respectively and $\alpha(f)$ is a frequency-dependent subtraction factor.

The formant tracks are predicted from the LP spectrum of speech obtained from LP spectral subtraction. The $k^{th}$ formant track $T_k(m) = [\hat{F}_k(m), \hat{F}_1(m-1), \hat{F}_1(m-2), ...]$ is estimated from the trajectory of the track up to time $m-1$ and the current formant observation $\phi(m)$ obtained from the estimate of LP spectrum of the current frame:

$$\hat{F}_k(m) = f(\phi(m), T_k(m-1))$$

(9)

where $f(\cdot)$ is an estimator. This classic estimation problem, of combination of a predicted trajectory with a noisy observation, can be approached using one of a number of solutions such as the Kalman filter. The prediction of a formant from its track can be formulated as

$$\hat{F}_k(m|m-1) = \sum_{i=1}^{P} c_{ki} \hat{F}_k(m-i)$$

(10)

Figure 4: Diagram of formant tracking in noisy speech
where \( c_{ki} \) are the coefficients of a lower order LP model of formant tracks.

The discrete-time Kalman filter\[ 6\] is adapted as follows.

\[
\hat{F}_k (m | m-1) = \sum_{i=1}^{p} c_{ki} \hat{F}_k (m-i) \tag{11}
\]

\[
P(m | m-1) = P(m-1) + Q \tag{12}
\]

\[
K(m) = P(m | m-1) (P(m | m-1) + R)^{-1} \tag{13}
\]

\[
\hat{F}_k (m) = \hat{F}_k (m | m-1) + K(m)(p_k (m) - \hat{F}_k (m | m-1)) \tag{14}
\]

\[
P(m) = (I - K(m))P(m | m-1) \tag{15}
\]

Where \( R \) is the measurement covariance matrix, updated by variance of differences between noisy observation and estimated tracks. The process matrix \( Q \) is set to 0.16 experimentally. The matrix \( P \) is initialized in the same way as \( Q \). Equation (11) and (12) are “prediction functions” and Equation (13), (14) and (15) are “measurement update functions”.

The whole of the formant tracking procedure is illustrated in Figure 4. In the formant candidate selection stage, formant candidates with large bandwidths are abandoned to eliminate the spurious formant candidates due to spectral subtraction. The initial estimated formant tracks are fed back to reclassify the formant candidates among the formant tracks according to the distance.

### 4.2 Performance Evaluation of Formant Tracking

The performance of the proposed method is evaluated using average track error measure. The results are shown in Table 4. Over 60% improvement through noise reduction has been achieved in the tracking the first formant, which is degraded most by the noise. In less affected higher formants (F2-F5), the improved method recovers the formant track with an average of 15% improvement. Figure 4 illustrates the example spectrogram of a clean speech superimposed with formant tracks of clean speech and recovered formant tracks of noisy speech.

### 5. DISCUSSION

This paper investigated the effects of car and train noise on formant estimation. Formant tracks from clean and noisy speech are obtained by 2D-HMM and LP analysis. Due to the characteristics of car and train noise, the first formant track is most affected by noise and the last formant is least affected. Correspondingly SNRs at the first formant frequencies in noisy speech are observed to be about 2dB to 5dB higher than overall SNRs for car and train noise respectively. This is also validated by the speaker-dependent speech recognition experiments, which shows in noise condition formant features achieved better performance than conventional MFCC.

A formant estimation method based on a combination of LP spectral subtraction and Kalman filter achieves some 60% improvement in the most affected first formant tracks. However current LP spectral subtraction does not take into account the continuity of the formants in that it is not fully integrated with track estimation. The discontinuity of formants is assumed to be compensated later by smoothing via Kalman filter. Hence, improvement of continuity of formant tracks in LP spectral subtraction and employment of better noise model are subjects for future research to achieve better formant track estimation in noisy condition, especially for high formants.

### 6. ACKNOWLEDGEMENTS

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

### 7. REFERENCE


<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR= 0</td>
<td>51.3</td>
<td>12.5</td>
<td>6.3</td>
<td>3.7</td>
<td>2.6</td>
</tr>
<tr>
<td>CLEANED</td>
<td>18.1</td>
<td>11.8</td>
<td>6.2</td>
<td>2.7</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 4: Average errors (%) of formant tracks in train noisy speech and cleaned speech.