ROBUST FORMANT TRACKING ALGORITHM IN REAL TIME NOISE ENVIRONMENT

Mandeep Singh Walia

Assistant Professor, Department of Electronics and Communication Engineering, UIET, Panjab University SSG Regional Centre (India)

ABSTRACT

The traditional approaches for formant frequency estimation are misled by spectral peaks in unvoiced speech and perform very poorly in transient background noise. Also, these traditional algorithms are not robust in real life noise environment and are unable to recover quickly after periods of silence. Robust formant tracking algorithm is discussed. Testing of the algorithm using various speech sentences has shown promising results over a wide range of signal to noise ratio's (SNR's) for various types of background included the babble (crowd of people), car, exhibition hall, restaurant, street, airport, train station and train recordings from different places. By observing the limitations of the traditional formant tracking algorithms it can be seen that the robust formant tracking algorithm is the most accurate algorithm. Results from algorithm showed that the robust formant tracking algorithm gives very good tracking performance in every environment.

Keywords*: AZF, *CEFS, DTF, LPC, Vocal Tract.

I. INTRODUCTION

The resonance frequencies of the vocal tract are called the formant frequencies or formants when vowels are pronounced. Formants can be found where there are large concentrations or peaks of energy in the spectrogram reading of a voiced sample. Accurate formant estimation for continuous speech (in real time noise environments) is a challenge because formant frequencies are not simple to track in such a dynamic environment. The formant estimation algorithm needs to be robust and be able to operate in a wide range of real-time noise scenarios. The vocal tract is generally considered as the speech production organ above the vocal folds. As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called formants. Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal. Voice verification systems typically use features derived only from the vocal tract. The human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the trachea (also called the "wind pipe") through the vocal folds (or the arytenoids cartilages). The excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these. Then model is made as like human speech production model. The robust formant tracking algorithm discussed in the present work is the most accurate formant tracking algorithm. This algorithm provides reliable formant frequency estimates for contrast enhanced frequency

shaping (CEFS) amplification (R. L. Miller, B. M. Calhoun, E. D. Young, 1999) and other applications. The Robust formant tracking algorithm (K. Mustafa and I. C. Bruce, 2006) is the most accurate algorithm for tracking formant frequencies of a speech signal. The paper is organized as follows. Section II gives an overview of traditional formant estimation techniques. Section III describes the proposed robust formant tracking algorithm. In section IV, experimental results are presented, and conclusions are given in section V.

II. TRADITIONAL FORMANT ESTIMATION TECHNIQUES

Development of accurate formant estimation algorithms began in the 1950s. Since then numerous techniques have been proposed for formant analysis. Most of the work can be classified as frequency domain techniques (such as picking peaks in the short-time frequency spectrum), parametric techniques (also called "analysis by synthesis") (R. W. Schafer and L. R. Rabiner, 1970) in which one generates a best match to the incoming signal based on a model of speech production.

III. TRADITIONAL FORMANT ESTIMATION TECHNIQUES

The traditional formant tracking algorithms do not track formants accurately. Robust formant tracking algorithm (K. Mustafa and I. C. Bruce, 2006), represent the best known formant analysis techniques and have been implemented in MATLAB. Fig. 1 shows a block diagram of the Robust Formant Tracker. A speech samples has been taken as an input signal for which formant estimation and tracking has to be done. After taking speech signal the first step in formant tracking with this algorithm is silence removal and pre-emphasis. Silence detection is usually based on the measuring some signal characteristics as relative energy level, zero crossing rate, first autocorrelation coefficient, first LPC linear predictor coefficient, first Mel – frequency cepstrum coefficient, and normalized prediction error. The easiest method proposed in this work to detect silent regions in speech is based on the computing of variations of the signal samples in speech frame, against the frame mean. If variations are big enough, the frame is considered as a speech frame, otherwise as a silence. Silent region is detected in the ways shown in Fig. 2 and Fig. 3. First, the mean of the frame samples is computed and then cumulative sum of absolute magnitude of differences between samples and mean is collected. Then if this sum exceeds predefined threshold the frame is considered as a speech frame, otherwise as a silent frame. Pre-emphasis is a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio by minimizing the adverse effects of such phenomena as attenuation differences or saturation of recording media in subsequent parts of the system. Voiced speech signals have a natural spectral tilt, with the lower frequencies (below 1 kHz) having greater energy than the higher frequencies. The lower frequencies have more energy because they contain the glottal waveform and the radiation load from the lips. In some speech processing applications it is desirable that this spectral tilt be removed by pre-emphasis or spectral equalization of the signal. A common method of pre-emphasis is to filter the speech signal using a High-Pass Filter (HPF) that attenuates the lower frequencies. Pre-emphasis is a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio by minimizing the adverse

effects of such phenomena as attenuation differences or saturation of recording media in subsequent parts of the system. Voiced speech signals have a natural spectral tilt, with the lower frequencies (below 1 kHz) having greater energy than the higher frequencies. The lower frequencies have more energy because they contain the glottal waveform and the radiation load from the lips. In some speech processing applications it is desirable that this spectral tilt be removed by pre-emphasis or spectral equalization of the signal. A common method of pre-emphasis is to filter the speech signal using a High-Pass Filter (HPF) that attenuates the lower frequencies. The result of the pre-emphasis is the approximate removal of the contribution of the glottal waveform and the radiation load effect from the lower frequencies of the signal, i.e. the energy in the speech signal is redistributed to be approximately equal in all frequency regions. Fig. 4 shows a spectrogram of a speech signal before and after it has been pre-emphasized using the filter from
Fig. 5.

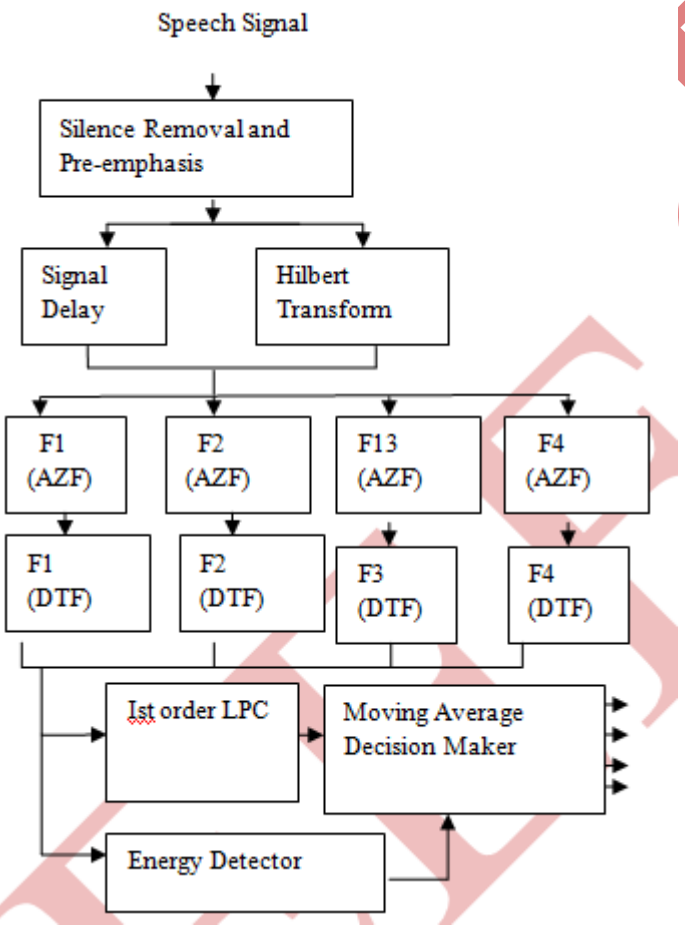

Fig. 1: Robust Formant Tracker

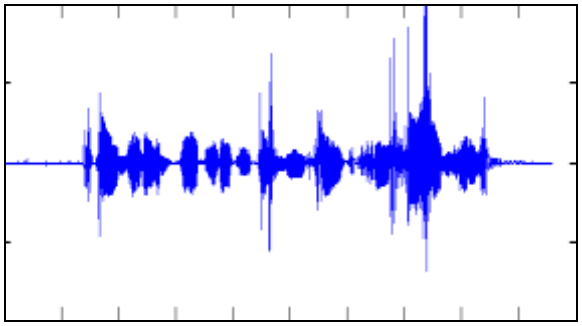

Fig. 2: Waveform Before Silence Removal

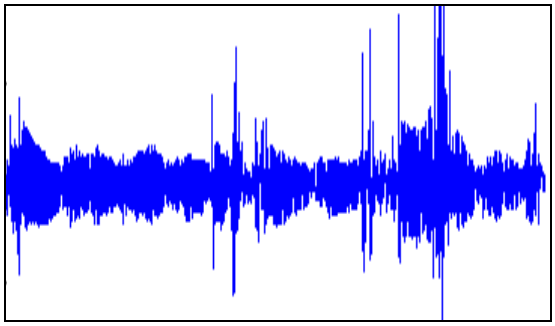

Fig. 3: Waveform After Silence Removal

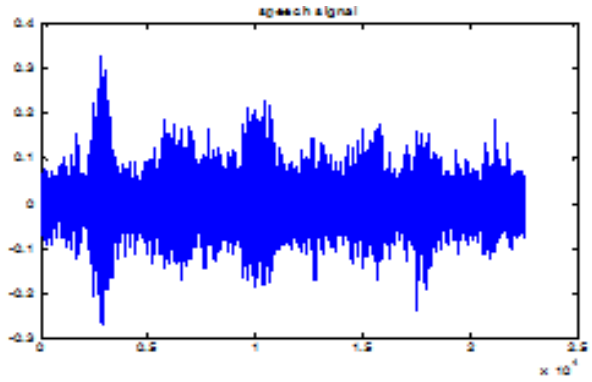

Fig. 4 Waveform before pre-emphasis

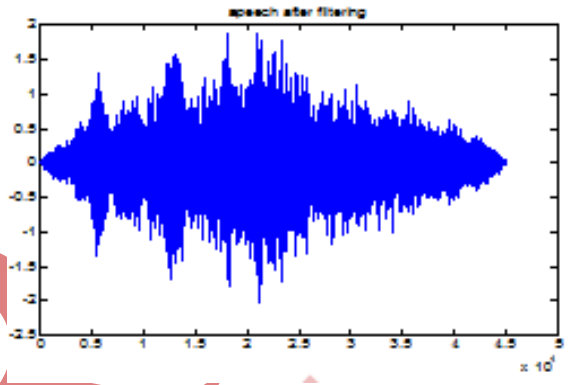

Fig. 5: Waveform after Pre-emphasis

Pre-emphasis is used due to voiced section of the speech signal usually falls at high frequencies due to high frequency formants have small amplitude compared to low frequency formants, and to reduce the DC effect. The frequency response of pre-emphasis is shown in Fig. 6.

After the signal has been pre-emphasized it is equalized to have a global RMS energy value of 0 dB. This equalization ensures that the energy threshold levels are set properly and to appropriate energy levels. An approximate, analytic version of the signal is then calculated to increase spectral accuracy for the formant estimates through an approximate Hilbert transformer (K. Mustafa and I. C. Bruce, 2006). The Hilbert transform of a function is defined as:

$$H\{f\}(y) = \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \frac{f(x)}{x-y} dx \quad (1)$$

Where PV stands for "principal value" and the primary reason behind converting the signal into its complex representation is to allow the use of complex filters in the formant filter bank (AZF's and DTF's). The real-time discrete signal, SR[n], can be represented by its complex form, SC[n], as:

$$SC[n] = SR[n] + jSH[n] \quad (2)$$

Where SH[n] is the Hilbert transform of SR[n]. The particular technique used to implement the Hilbert transformer in the formant tracking algorithm uses an optimum FIR filter. The Hilbert transformer is implemented with a 20th-order linear-phase FIR filter designed using the Parks-

McClellan algorithm (Remez exchange algorithm). The frequency and phase responses of the filter are shown in Figure 7. The filter is designed using the Remez exchange algorithm and Chebyshev approximation to have an optimal fit between the desired and actual frequency responses. The analytic signal is then filtered into four different bands using a bank of adaptive band-pass filters (called Formant filters). Each of the four formant filters (F1, F2, F3 and F4) in the filter bank is made up of an All- Zero Filter (AZF) and a Dynamic Tracking Filter (DTF) (A. Rao and R. Kumaresan, 2000).

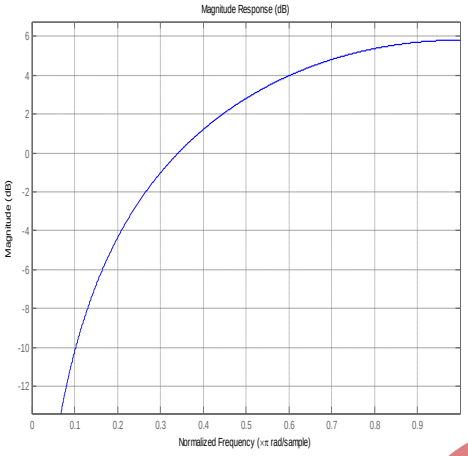

Fig. 6: Frequency Response of Pre-emphasis

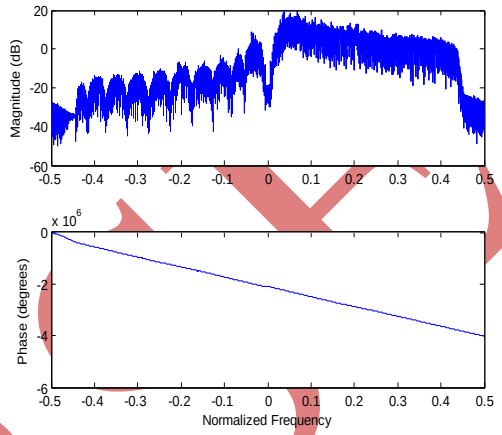

Fig. 7: Frequency Response and Phase Response of the Filter

The zeros of each of the AZF's are set to the latest estimate of the formant frequencies from the other three bands. The DTF provide the single pole located at the latest estimate of the formant frequency for that band. This cascade arrangement results in each of the filters having a pole around its own formant frequency and zeros at the other formant frequency locations. Each of the four band-pass filters allows only the signal around the frequency region of the desired formant to pass and suppresses the other frequency regions. The formant filter bank has a fundamental modification that the F1 filter of the filter bank has an added zero at the pitch frequency (F0) for further suppression of the region below the F1 frequency (the pitch region). This decreases the effects of the pitch on the F1 estimate.

The adaptive band-pass filter bank used in the formant tracking algorithm is similar to the one proposed (A. Rao and R. Kumaresan, 2000) but it has a modified first formant filter that removes the effects of the pitch from the first formant filter band. Each channel of the filter bank consists of an all-zero filter (AZF) cascaded with a single pole dynamic tracking filter (DTF). The combination of the AZF and the TF is called a formant filter (A. Rao and R. Kumaresan, 2000) and is responsible for tracking one individual formant frequency. The filters are designed in the complex domain because it is easier to design the unity gain and zero phase lag filters in the complex domain. Adaptively varying the zeros and pole of each formant filter, allows the suppression of interference from neighboring formant frequencies and from other spectral noise sources, while tracking an individual formant frequency as it varies with time.

The transfer function of the kth AZF at time sample index n is:

$$H_{AZFK}(n,z) = K_k[n] \; X \prod_{\substack{l=1 \\ l\neq k}}^{4} (1 - r_z e^{2\pi f_l[n-1]} z^{-1} \quad (3)$$

Where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1 \\ l\neq k}}^{4} (1 - r_z e^{2\pi (f_l[n-1] - f_k[n-1])}} \quad (4)$$

and rz is the radius of the zeros on the Z-plane, fl[n-1] is the formant frequency of the l^{th} filter estimated at time index n–1 and, fk[n] is the formant frequency of this filter (k^{th} filter) estimated at index n–1. The gain of $K_k[n]$ ensures that the AZF has unity gain and zero phase lag at the estimated formant frequency of the k^{th} component. A wide range of values for rz were tested and the best results were obtained (for the range of values tested) for rz = 0.98.

The 'DTF' in each formant filter is a single-pole dynamic tracking filter. The pole location is always set to the previous estimate of the formant frequency of that formant filter. The transfer function of the kth DTF at index n is:

$$H_{DTFk}(n,z) = \frac{1 - r_p}{(1 - r_p e^{j2\pi fk[n-1]} z^{-1})} \quad (5)$$

Where r_p is the radius of the pole and $f_k[n-1]$ is the formant frequency of the k^{th} filter at time index n-1. A wide range of values for r_p were tested and the best results were obtained (for the range of values tested) using r_p = 0.90. The transfer function of the 1st formant AZF is slightly different than that of the other AZFs. The AZF of the first formant filter has an additional zero at the location of the pitch estimate to suppress pitch effects on the first formant estimate. The transfer function of the 1st AZF at index n is:

$$H_{AZF1}(n,z) = K_k[n] X \prod_{\substack{l=1 \\ l\neq k}}^{4} (1 - r_z e^{j2\pi (f_l[n-1]} z^{-1}) \quad (6)$$

Where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1 \\ l\neq k}}^{4} (1 - r_z e^{2\pi (f_l[n-1] - f_k[n-1])}} \quad (7)$$

and f0 [n-1] is the pitch estimate at time index n-1, that is provided to the 1st formant filter by the gender detector (K. Mustafa and I. C. Bruce, 2006). After the placement of the pole and zeros for each of the formant filters, the transfer function and the complex filter coefficients of the four formant filters are calculated. These complex filter coefficients are then used to filter the analytic speech signal into four band-limited spectral regions from which the four formant frequencies are estimated.

The frequency responses of the four formant filters are set as; pitch (F0) is set to 200 Hz, the first formant frequency (F1) is set to 700 Hz, the second formant frequency (F2) is set to 1500 Hz, the third formant frequency (F3) is set to 2200 Hz and the fourth formant frequency (F4) is set to 3500 Hz. After the speech signal has been filtered using the adaptive band-pass filterbank of the signal over the previous 20 ms in each band is calculated. In order for the algorithm to estimate a particular formant frequency from the spectrum (instead of using the moving average value), the energy calculated in that formant band has to be above a certain 'energy threshold level', in addition to that speech segment being voiced. Linear prediction provides a good model of the speech signal. This is especially true for the quasi steady state

voiced region of speech in which the all pole model of LPC (Kaneko, Takuma, and Tetsuya Shimamura, 2014) provides a good approximation to the vocal tract spectral envelope. During unvoiced transient region of speech, the LPC model is less effective than for voiced regions, but it still provides acceptably useful models. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time varying system. Linear prediction models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. For vowel sounds and other voiced regions of speech, which have a resonant structure and high degree of similarity overtime shifts that are multiples of their pitch period, this modeling produces an efficient representation of the sound.
The linear prediction problem can be stated as finding the coefficients a_k which results in the best prediction (which minimizes mean-squared prediction error) of the speech sample s [n] in terms of the past samples as s[n-k], k = {1… P}.The idea behind linear prediction is to approximate each sample of the speech signal as a linear combination of past samples. By minimizing sum of squared differences between the actual speech samples and the predicted ones, a unique set of predicted coefficients can be determined. A linear predictor of order p is defined as:

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \quad (8)$$

Where $\tilde{s}[n]$ is the prediction of s[n] by the sum of p past weighted samples of s[n]. The system function of the p^{th} order predictor is a FIR filter of length p given by:

$$p(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \quad (9)$$

The associated prediction error filter is:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = 1 - p(z) \quad (10)$$

And prediction error is defined by:

$$e[n] = s[n] - \tilde{s}[n]$$

$$= s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \quad (11)$$

The roots of the inverse of the prediction error filter corresponds to the poles placed to model the original signal as closely as possible while minimizing the mean squared error between the estimated and original signals. First order linear prediction (p = 1) obtains one linear predictive coefficient and the corresponding single pole is placed to model the original signal as well as possible. Second order LPC tries to model the original signal using two poles, and so on. The first four formant frequencies of the speech signal are estimated from the four filter bands of the adaptive band pass filter bank using first-order LPC. The analytic signal from each of the bands is first windowed using a 20-ms periodic Hamming window and then the linear predictive coefficient (one per band) of the previous 20 ms of the windowed signal from each band is calculated. LPC tries to fit a single pole model to each signal and the location of the pole corresponds roughly to the vocal tract pole (formant frequency) in that band, for voiced segments of speech. The LPCs are only calculated from the bands if the entire previous 20-ms window of the speech signal is voiced. A first-order Linear Prediction Coefficient (LPC) (Sahoo, D. K.,

et al., 2013) is then calculated for the analytic signal in each of the four bands. From each of these coefficients a formant frequency estimate is obtained. As the value of the four formant frequencies vary with time, the formant pre-filters are modified to track them by changing their pole and zero locations. Due to the band-pass pre-filtering of each formant frequency region prior to LPC, the frequency estimates provided by LPC are more accurate and the algorithm is less susceptible to errors due to background noise. The formant estimation is further refined by adding an adaptive voicing detector to detect a voiced and unvoiced speech segments. LPC estimates for the formant frequencies are only used during the voiced segments of speech. During the unvoiced speech segments or when the signal energy of a particular formant frequency region (determined by the adaptive energy detector) is below a set threshold level, the formant frequency estimates are assigned their moving average value. This approach ensures that the formant tracker is able to recover quickly and with minimum error to the formant estimates, after unvoiced or low-energy speech segments. The energy detector threshold levels are also made adaptive for each of the formant filters so that they can adjust to long term changes in the energy levels of each formant frequency region. The voicing detector calculates the log ratio between the energy in the lower and higher frequencies of the speech to determine if a speech segment is voiced or unvoiced. If there is more energy in the lower frequencies than the higher ones, the speech segment is classified as being voiced. The voicing detector also has a threshold with hysteresis to ensure that switching from voiced to unvoiced speech (or vice versa) does not erroneously occur too quickly. Finally, an autocorrelation-based energy test is performed to ensure that voicing is not detected erroneously when there is no actual voicing in the speech but sufficient energy is present in the lower frequencies due to 'colored Gaussian noise' (or other background noises) . The voicing detector provides a sample by sample decision on whether a segment is voiced or unvoiced. In order for the voicing detector to work properly for both male and female speakers, various parameters of the voicing detector need to be modified. The main purpose of the gender detector is to determine the gender of the speaker and pass this information to the voicing detector so that it is able to modify its parameters. The gender detector uses a pitch based method to classify the gender of the speaker where the pitch is calculated using an autocorrelation based method. The gender detector also provides the pitch estimate to the first formant filter so that an additional zero can be added at the location of the pitch in the AZF of the first formant filter. Extensive testing of the robust formant tracking algorithm has been done which showed that the formant tracking algorithm is robust to a wide variety of real-time background noise conditions. The algorithm is able to provide reliable formant frequency estimates from continuous speech for both male and female speakers. It recovers quickly and with minimal error when problems do occur and when there is a switch in speakers. The moving average decision maker assigns the estimated value to the formant frequencies (from the LPCs) only when the segment is voiced and the energy of the formant frequency is above its respective threshold level. If the segment is unvoiced or if the energy of a particular formant is below its respective threshold level, then the current value of the formant frequency decays toward the moving average value for that formant frequency according to:

$$F_i[n] = F_i[n-1] - (0.002 * (F_i[n-1] - F_i^{MA}[n-1])) \quad (12)$$

Where F_i is the formant estimate of i[th] formant frequency at time index n and $F_i^{MA}[n-1]$ *i* is the previous value of the moving average for the i[th] formant frequency. The update rule for the moving average value of each formant frequency:

$$F_i^{MA}[n] = \frac{1}{n} F_i[k] \quad (13)$$

Where F_i^{MA} the moving average is value for the i[th] formant frequency at index n and $F_i[n]$ is the estimate of the i[th] formant frequency at index n.

IV. EXPERIMENTAL RESULTS

The primary goal of formant tracking algorithms is to develop a reliable formant tracking algorithm that is robust in real-time noise scenarios. Different test cases are described and the performance of the algorithms under these conditions has been discussed. The formant tracking algorithms has been tested using a noisy speech corpus (NOIZEUS) (ITU-T P.56, 1993). The noisy database contains IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database (H. Hirsch, and D. Pearce, 2000) and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport, and train-station noise. The noise signals were added to the speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB. This corpus is available free of charge. Algorithm is tested for both male and female voices. The spectrogram of male speaker saying "Her purse was full of useless trash is shown in fig. 8. The spectrogram of female speaker saying "The clothes dried on a thin wooden rack" shown in fig. 9.

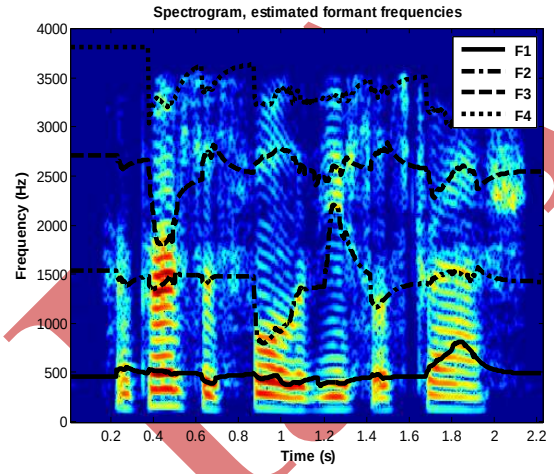

Fig. 8: Spectrogram and Formant Frequencies for a Male Speaker Saying "Her Purse was full of Useless Trash"

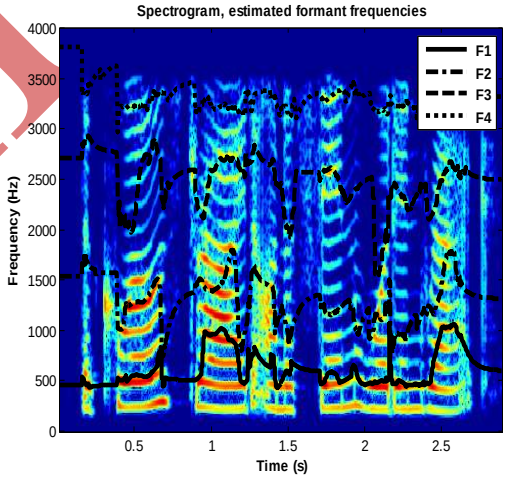

Fig. 9: Spectrogram and Formant Frequencies for a Female Speaker Saying "The Clothes Dried on a Thin Wooden Rack"

The IRS filter is independently applied to the clean and noise signals. The active speech level of the filtered clean speech signal is first determined using the method B of ITU-T P.56 (ITU-T P.56, 1993). Table 1 shows the range of first four formant frequencies. Generally the estimated formant frequencies vary in these ranges.

TABLE 1: Ranges for Formant Frequencies

Formants	Minimum Frequency (Hz)	Maximum Frequency (Hz)
F1	270	730
F2	840	2290
F3	1690	3010
F4	2500	4500

V. CONCLUSION

Robust formant tracking algorithm has been discussed in this work. The present work shows the testing of robust tracking algorithm for a speech signal in different environmental conditions. The operation of the algorithm is tested and analyzed in the presence of background noises. Quantitative analysis has shown that it provides accurate formant frequency estimates for both male and female speakers for a wide range of SNR's in real-time noise conditions. The robust formant tracking algorithm recovers quickly after erroneous estimates to go back to tracking the actual formant frequencies in the speech signal. The algorithm occasionally gives choppy and oscillating formant frequency estimates. This is an undesirable result because the actual formant frequencies of speech normally vary slowly with time and have smooth transitions. Furthermore, the estimated formant frequencies have to be smooth and the algorithm has to be able to identify formant transitions accurately. The oscillating formant frequency problem may be solved in future updates to the formant tracking algorithms by either smoothing the formant frequency estimates or by incorporating additional logical limitations to prevent abnormal jumps in the formant estimates. Another future improvement may be to modify the formant pre-filters to have variable bandwidths that are dependent on the magnitudes of the poles estimated by the linear prediction coefficients.

REFERENCES

[1] Miller, R. L., Calhoun, B. M., and Young, E. D., Contrast enhancement improves the representation of /E/-like vowels in the hearing-impaired auditory nerve, J. Acoust.Soc. Am., 106, 1999, 2693–2708.

[2] Mustafa, K., and Bruce I. C., Robust formant tracking for continuous speech with speaker variability, IEEE Transactions on Audio, Speech and Language Processing 14(2), 2006, 435–444.

[3] Schafer, R. W., and Rabiner, L. R., System for automatic formant analysis of voiced speech, J. Acoust. Soc. Am., 47, 1970, 634–648.

[4] Rao, A., and Kumaresan, R., on decomposing speech into modulated components, IEEE Trans. Speech Audio Processing, 8, 2000, 240–254.

[5] Kaneko, Takuma, and Tetsuya Shimamura, Noise-Reduced complex LPC analysis for formant estimation of noisy speech, 2014

[6] Sahoo, D. K., et al, Estimation of formant frequency of speech signal by linear prediction method and wavelet transform, International Journal of Engineering Research and Technology, 2(3), 2013.

[7] ITU-T P.56, Objective measurement of active speech level, 1993.

[8] Hirsch, H., and Pearce, D., The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions, ISCA ITRW ASR, Paris, France, 2000, 18-20.

Biographical Notes

Mandeep Singh Walia is working as an Assistant Professor in Electronics and Communication Engineering Department, PUSSGRC, Hoshiarpur.