Improving Speech Intelligibility in Cochlear Implants using Acoustic Models

P. VIJAYALAKSHMI[1], T. NAGARAJAN[2] and PREETHI MAHADEVAN[1]
[1]Department of Electronics and Communication Engineering,
[2]Department of Information Technology
SSN College of Engineering, Chennai, Tamil Nadu, India
vijayalakshmip@ssn.edu.in, nagarajant@ssn.edu.in

Abstract: - Cochlear implant (CI) is a prosthetic device that partially replaces the functions of the human ear via electrical stimulation. Cochlear implants are system and/or patient specific that mandates a simulation model prior to implantation. In the present work to improve the perceptual quality of the speech generated by a CI model, system specific parameters are analyzed by developing uniform bandwidth filterbank-based acoustic CI models, an auditory model-based CI system with frequency bands spacing similar to the critical-bands of an auditory system and Mel-frequency cepstral coefficients (MFCC) based analysis-synthesis system for cochlear implants. Acoustic CI simulations are generated for all the vowels of English language and words (easy and hard) from the Lexical Neighbourhood Test (LNT) and sentences from TIMIT database using waveform and feature extraction strategies. A closed-set listening test is conducted and a comparative study is made among the various acoustic CI models developed. The perceptual quality/speech intelligibility of the speech is rated in 5 point grading. It is observed that the acoustic CI simulation for sentences generated by critical-band-based CI system showed a mean opinion score of 4.1 as opposed to 3.1 for uniform bandwidth filters-based CI system.

Key-Words: - Cochlear implants, Filter banks, Critical band, Speech intelligibility, MFCC, Channel vocoder, Auditory model, Acoustic CI simulations.

1 Introduction

Cochlear implants (CI) are prosthetic devices that partially restore hearing via electrical stimulation of the inner ear in individuals with severe to profound hearing loss [1], [2]. Cochlear implant bypasses the normal hearing operation by directly stimulating the auditory nerve through the electrodes. The main function of this prosthetic device is an artificial replacement of damaged inner ear using an (i) external body worn speech processor and an (ii) internal receiver-stimulator [3].

A microphone picks up the speech signal; the external speech processor processes the speech signal using various speech processing strategies, and generates an encoded speech data to be received by the internal receiver stimulator through RF (radio frequency) link. The receiver decodes the signal and transmits the specified stimulation waveform to the intra-cochlear electrodes [1], [2]. The receiver stimulator is used for stimulating the auditory nerve via electrode array that enables understanding of the human speech by brain.

Cochlear implants are developed to improve the listening capability of people whose hair cells in the cochlea are non-functional [4]. According to [5] the speech intelligibility or the perceptual quality of the speech generated by the cochlear implant devices has reached only 70% of that of the normal hearing capability that can be attributed to many of the factors, one of which is speech signal processing. Speech processing strategies have an important role in maximizing the user's communicative potential.

Design of cochlear implants started with single channel implants. These implants provided electrical stimulation at a single site in the cochlea using single electrode and tested with patients in the early 1970s that were capable of conveying time/envelope information, limited spectral information and had insufficient speech recognition [1], [6]. In order to improve the spectral information, multi-channel cochlear implants are developed in 1980s, which has multiple sites in the cochlea to provide electrical stimulation using an array of electrodes.

Multi-channel cochlear implants are the widely used systems. The major speech processing strategies that are used in the multichannel cochlear implants are waveform and feature extraction strategies based on whether the speech waveform or the feature is processed. The compressed analog (CA) approach and continuous interleaved sampling (CIS) approaches [7] are the variations of the waveform strategy based on simultaneous or non-simultaneous stimulation of the electrodes. The SpectralPEAK (SPEAK) and MultiPEAK (MPEAK) are some of the variations in the feature extraction strategy. The SPEAK strategy encodes the spectral maximum information (formants), where the incoming signal is passed through a set of 20 bank of filters with center frequencies ranging from 250 Hz to 10 kHz and only six to ten filter outputs with maximum amplitude (N-of-M) are selected [1], [2], [6], [8]. In MPEAK strategy high frequency information are also extracted apart from the first four formant frequencies [1].

Cochlear implants with various speech processing strategies as described above restores hearing to a greater extent. However, there are considerable variations in the CI patient outcomes [9]. This variability in the cochlear implants is patient specific and/or system specific. Therefore the patient/system specific parameters have to be optimized for individual patients. Cochlear implants are surgical procedures that mandate the estimation of the effect of these parameters on perceptual quality of the processed speech before patient evaluation. Towards this end, in the current work various acoustic CI simulation models are generated to study the effect of system specific parameters such as, number of channels, type of filters, bandwidth requirement of each of the filter in the filter bank, etc. The waveform and feature extraction strategies are implemented and to further improve the speech intelligibility experiments are performed with the bank of filters based on the structure of critical-bands of the cochlea and feature-based analysis-synthesis system.

This paper is organized as follows: Details of the speech corpora used for the current work is described in the following section. Section 3 describes the basic CI simulation model developed for the present work. Section 4 describes the uniform bandwidth filter-bank-based acoustic CI model and the perceptual test conducted for vowels, words, sentences, etc. In Section 4 for vowel estimation formant filter-based acoustic CI simulation models are also developed and compared with the performance of uniform-bandwidth filterbank based system. To improve the perceptual quality of the speech auditory-model based CI system is developed and evaluated as described in Section 5. In Section 6 a feature-extraction based speech processing strategy based on Mel-frequency cepstral coefficients (MFCC) is implemented and is validated using closed-set listening test.

2 Speech Corpora

To study the effect of variations in the system parameters on the speech intelligibility speech utterances such as vowels, words and sentences are collected from normal speakers. Vowels such as /a/, /i/, /e/, /o/ and /u/ are collected from 10 normal male and 10 female speakers. The easy and hard words (a total of 150 words) from list-1 of lexical neighborhood test (LNT) [10] are collected from a male and a female speaker. The speech data analyzed in this study are recorded using a head mounted microphone whose frequency response is 20 Hz to 20 kHz and sampled at a rate of 16 kHz. For the analysis on sentences the TIMIT speech corpus [11] is used. TIMIT database contains a total of 6300 sentences uttered by 630 male and female speakers from 8 major dialects of American English, each reading phonetically rich sentences.

3 The basic CI simulation model

The latest and most successful signal-processing strategies used in cochlear implants are based on vocoder principles [12]. Therefore the CI simulation models developed for the current work are based on the analysis-synthesis system of channel vocoder. The channel vocoder consists of a speech analyzer and a speech synthesizer. The analysis section consists of a microphone to pick up the speech signal and a speech processor to decompose the signal into its frequency components. Speech processor consists of a bank-of-filters (band-pass filters) in order to filter the incoming speech signal into a number of contiguous frequency channels. The envelope of the signal in each channel is estimated by full-wave rectification and low-pass filtering. In addition to envelope estimation, the vocoder analyzer makes a voiced/unvoiced decision and estimates the pitch frequency. The synthesizer modulates the received envelopes by the appropriate excitation as determined by the voiced/unvoiced decision [13], [14], [15].

The excitation signal for voiced speech segments consists of train of impulses separated with the pitch period controlled by pitch frequency as determined in the analyzer section. For the unvoiced speech segments random noise acts as an excitation signal. The modulated signals are subsequently band-pass-filtered by the same filters as used in the analyzer and then added together to produce the synthesized speech waveform.

For the current work to analyze the system specific parameters such as type of filters, numbers of channels, band of frequency, etc. various cochlear implant simulation models are developed. Initially a uniform-bandwidth filters-based acoustic CI model is developed using FIR and IIR (Chebyshev, Butterworth and Bessel) filters and their filtering effects on the speech signal are analyzed. Based on the performance of these filters in the roll off region, stability of the filters, order of the filters, etc., it is observed that Chebyshev type-2 filter with order 5 with stop-band attenuation of 40 dB is found to be a suitable filter. For all our further analysis a bank of Chebyshev filter with the specification mentioned above is used.

4 Uniform-bandwidth filters-based acoustic CI simulation

As discussed in the previous section channel vocoder based analyzer is developed using a bank of Chebyshev type-2 uniform-bandwidth band-pass filters over a frequency range of 200 Hz to 7000 Hz with a non-overlapping uniform bandwidth of 400 Hz apart from a low-pass filter with a cut-off frequency of 200 Hz, making a total to 18 filters in the bank [16]. The specifications of the filters used are maintained for both the analysis and synthesis sections of the CI model. Using the CI model developed acoustic CI simulations are generated for the speech data collected as described in section 2.

4.1 Vowel estimation

In the analysis section, the speech signal corresponding to each of the vowel is passed through the bank of 18 uniform bandwidth filters. Envelope estimation of each of the filter output is performed by passing the filtered signal through full-wave rectifier followed by a low-pass filter with a cut-off frequency of 200 Hz. Fig 1(b) shows the output of the channel 5 of bandwidth 400 Hz with the lower and upper cut-off frequency of 1400 Hz and 1800 Hz respectively and the corresponding full-wave rectified output and the envelope of the filtered output are illustrated in Fig. 1(c) and 1(d) respectively.

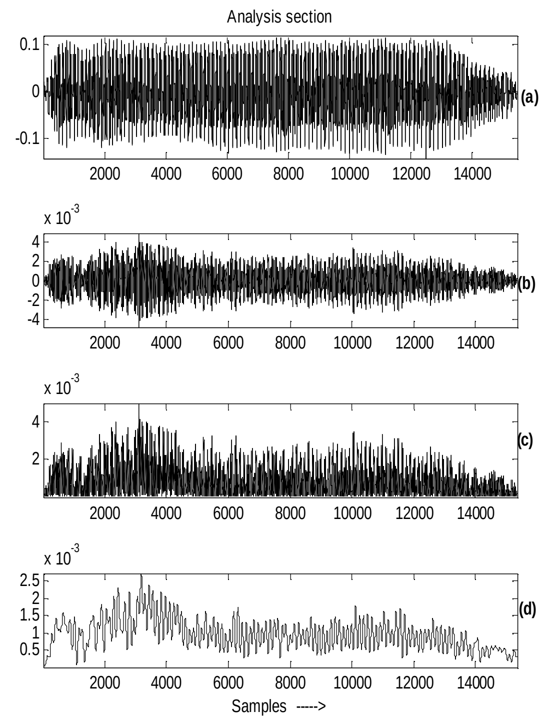

Fig 1.Uniform-bandwidth filter-based CI system - the analysis-section (voiced speech) (a) speech signal for the vowel /a/, (b) filtered output of channel 5 (c) full-wave rectified output of channel 5 and (d) the corresponding envelope.

As all vowels are voiced signals voiced/unvoiced classification need not be performed for vowel estimation. Pitch period of each of the vowels uttered by male and female speakers are estimated using the simplified inverse filter tracking (SIFT) [17]. This technique involves inverse filtering the linear-prediction based magnitude spectrum and extracting the corresponding time domain signal which is the residue containing the pitch information. Autocorrelation of the residual signal provides the estimation of the pitch period as shown in Figs. 2 and 3. It is observed that for female speakers the pitch period varies between 3 and 5 ms and for male speakers between 6 and 9 ms.

The analyzer outputs such as the envelopes estimated for each of the filter in the filter-bank, pitch period etc., are passed to the synthesizer section. As discussed in section 3 the envelopes are modulated using train of impulses separated by pitch period as estimated in the analyzer section as shown in Fig 4(a). Fig. 4(b) shows the filter output of channel 5 at the synthesis section.

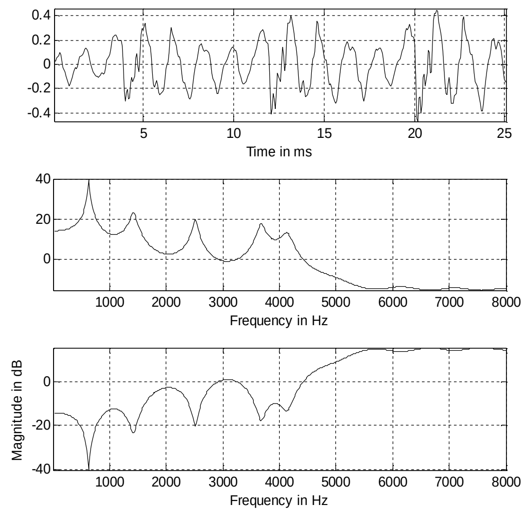

Fig 2. Simplified inverse filtering technique showing a frame of speech of vowel /a/ followed by LP spectrum and inverse spectrum

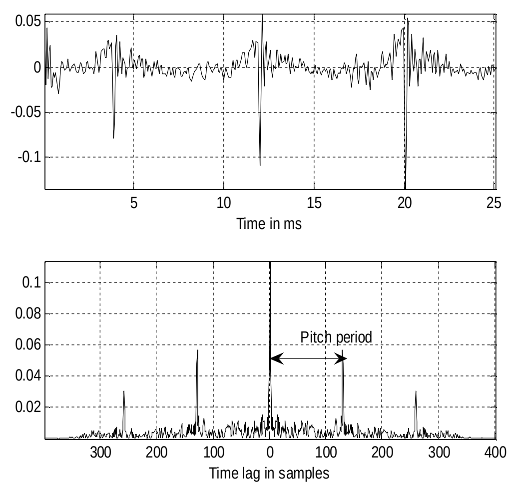

Fig 3.Simplified inverse filtering technique showing residue and the corresponding autocorrelation function denoting the pitch period in samples.

The modulated signals from each of the channels are subsequently filtered through the same bank-of-filters as used in the analyzer section and the resultant signals are summed up to obtain the synthetic speech as shown in Figs 4(c) and 5. Synthetic speech signals are generated for all the male and female speakers' speech data (vowels) as described in Section 2.

One of the major issues in the implementation of CI models is the number of channels (filters) required for the filter bank. That is, the determination of the bands-of-frequencies those are

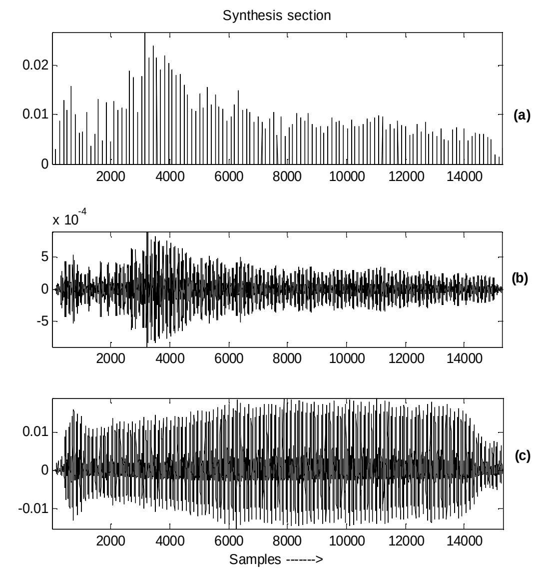

Fig 4. Uniform-bandwidth filter - synthesis section (a) Envelope modulation with train of impulses, (b) corresponding filtered output and (c) synthetic speech corresponding to vowel /a/.

.

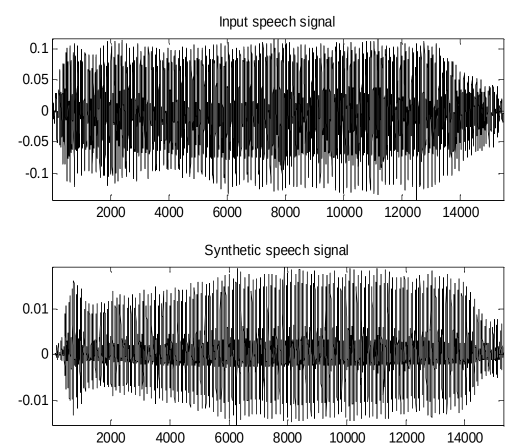

Fig 5.Uniform-bandwidth filter-bank based CI-model - Input and Synthetic speech signal for the vowel /a/.

more influential than the others. This is carried out by designing non-overlapping contiguous filters with various uniform-bandwidths thereby varying the number of filters in the filter bank. For the current work the numbers of channels are varied from 6-18 in steps of 3 with bandwidths of the filter banks varying uniformly from 1200 Hz to 400 Hz apart from the low-pass filter with a cut-off frequency of 200 Hz. Synthetic speech signals are generated for all the vowels uttered by all the male and female speakers.

4.2 Formant filter-based vowel estimation

Based on the spectral content of the given speech signal, in a uniform-bandwidth filterbank analysis described in the previous section, the estimated envelopes in the analyzer showed a wide variation in the magnitude. The relative magnitude of the estimated envelopes gives an insight in reducing the number of channels by choosing only channels with higher magnitude alone. That is, this observation leads to a fact that instead of having bank-of-filters over the entire range of frequencies of the speech signal filters can be designed only for spectral maxima, the formant frequencies. This is carried out by estimating the first four formant frequencies of all the vowels uttered by both male and female speakers.

Formant frequencies are estimated using linear prediction-based formant extraction technique [18]. First four formant frequencies are extracted in a frame-by-frame analysis of a given speech signal with a frame size of 25ms, thereby reducing the number of channels, in the analysis and synthesis sections of the channel vocoder, to four. The cut-off frequencies of these filters, for a given vowel, are determined from the statistics (mean, min. and max) derived from the formant frequencies extracted for all the speakers. Synthetic speech signals are generated with the procedure as described in section 4.1. Fig. 6 shows the corresponding input and synthetic speech signal for the vowel /a/.

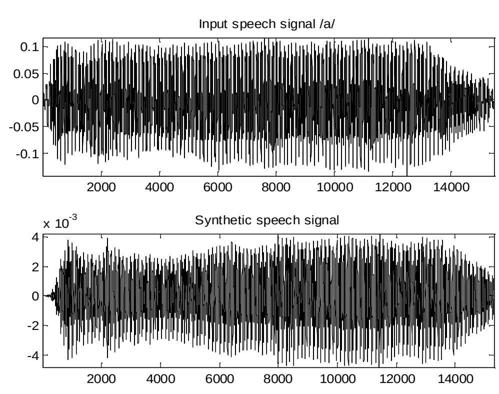

Fig 6.Formant synthesis - Input and Synthetic speech signal for the vowel /a/.

4.3 Word and sentence estimation

As opposed to vowels, words and sentences uttered by speakers contain both voiced as well as unvoiced speech segments. This mandates voiced/ unvoiced/silence (VUS) classification in the vocoder analyzer. For the current work the VUS discrimination is carried out by building a 3-mixture component Gaussian mixture model (GMM) [19].

4.3.1 Feature extraction

The features considered for voiced/unvoiced/silence classification are (a) normalized short-term energy, (b) short-term zero crossing-rate and (c) main-lobe width of short-term autocorrelation function. The amplitude of the unvoiced segments in speech is generally much lower than the amplitude of the voiced segments [14]. The short-term energy of the speech signal provides a convenient representation that reflects these amplitude variations. Mathematical representation of short-time energy, E_n, is given by,

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad \dots\dots\dots\dots (1)$$

Owing to the excitation signals from which the voiced and unvoiced speech segments are generated, most of the energy of a voiced segment is concentrated in the low frequency region whereas for an unvoiced speech segment in the high frequency region leading to low and high zero crossing rates respectively [14]. Short-time zero crossing rate, Z_n, is mathematically represented by,

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|.w(n-m) \quad \dots\dots\dots\dots (2)$$

Short-term zero crossing rate and short-time energy are calculated for all voiced and unvoiced speech segments from vowels and words. The third feature that discriminates voiced from unvoiced is the autocorrelation function which is given mathematically by,

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m).w(n-m).x(m+k).w(n-k-m) \quad \dots\dots\dots\dots (3)$$

The inherent quasi-periodic nature of the voiced speech signal causes a wide main lobe width of the autocorrelation function. For an unvoiced speech segment lack of periodicity causes a narrow width of the main lobe (refer to Fig. 7).

The three mixture components in the GMM correspond to silence, voiced and unvoiced speech. Since it is 3-class problem, from the mean vectors of the mixture components, one can trivially find the mapping between the 3-classes (voiced, unvoiced, and silence) and the mixture components. Here, the GMM is used to tokenize the sequence of feature vectors. Figs. 8 and 9 describe the distribution of zero crossing rates for a pair of classes (voiced & silence, and unvoiced & silence respectively).

For word estimation the steps involved are similar to that of the vowel estimation in the analyzer as described in section 4.1 except for the fact that voiced/unvoiced/silence classification has to be carried out for the given speech signal. In precise, the synthesis section receives the following data from the analyzer; (i) envelopes for each of the channels in the filter-bank, (ii) voiced/unvoiced/ silence classification, and (iii) pitch frequency as estimated using inverse filtering technique for the voiced segments. As discussed in section 3 the envelopes corresponding to the voiced speech are

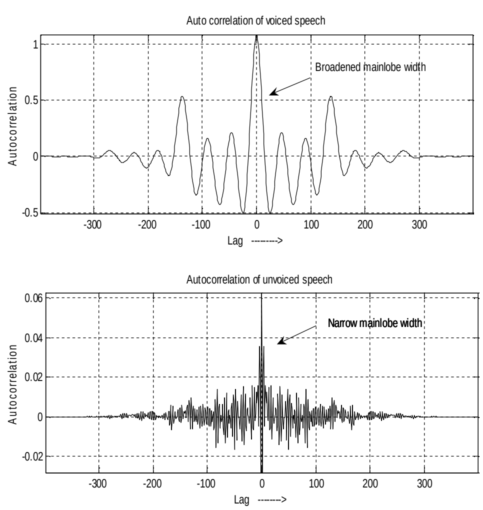

Fig. 7 Autocorrelation function of voiced and unvoiced speech segment showing varying main-lobe widths

modulated using train of impulses separated by pitch period. Whereas modulation of unvoiced speech segments is carried out by the excitation signal generated from a white noise generator. The modulated voiced and unvoiced speech segments are subsequently filtered by the same set of filters as that of the analyzer and summed up to obtain the synthetic speech signal. Using the above procedure word estimation is carried out for all the 150 words uttered by a male and female speaker each.

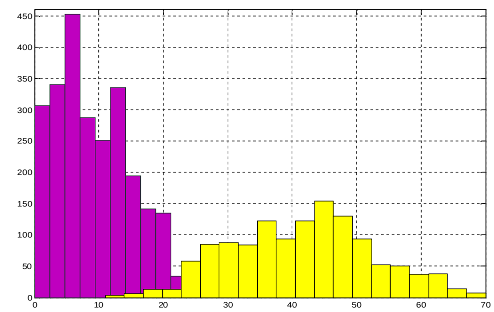

Fig. 8 Distribution of zero-crossing rate for voiced and silence segments.

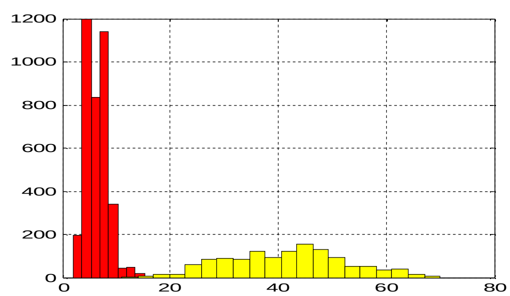

Fig. 9 Distribution of zero-crossing rate for unvoiced and silence segments.

Figs. 10 and 11 show the outputs of analysis and synthesis section for the word 'sugar' from the LNT data uttered by a male speaker using uniform-bandwidth filters-based CI model. Fig. 10(b) shows the channel 5 output of unvoiced region in the word ‘sugar’. Figs.10(c) and (d) show the full-wave rectified output and the corresponding envelope. Figs 11(a) and (b) illustrate unvoiced speech segment modulated with white noise and the corresponding filter output at the synthesis section. For comparison the input speech signal and synthetic speech signal for the word 'sugar' is shown in Fig 12.

Similarly sentence estimation is performed for 380 sentences spoken by 38 male and female speakers from dialect-1 of TIMIT training data. Fig. 13 shows an example for a sentence, uttered by a female speaker, estimated using uniform bandwidth

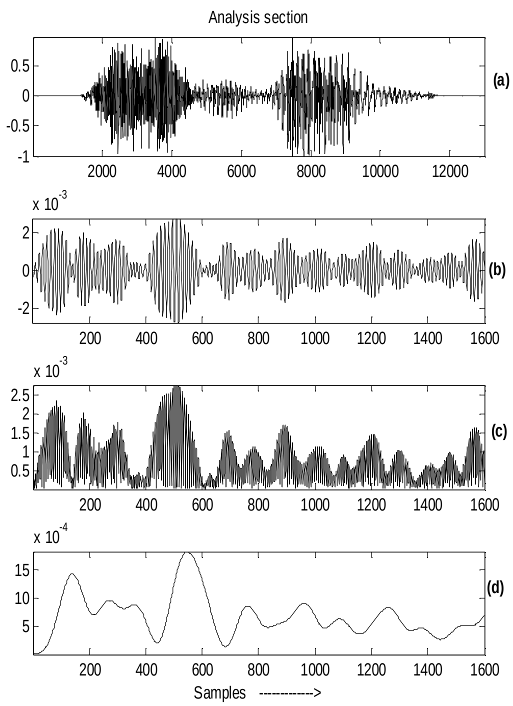

Fig 10. Uniform-bandwidth filters-based CI system - Analysis section (a) Input speech signal for the word 'sugar' (b) channel 5 output for the unvoiced region (c) corresponding full-wave rectified output and (d) envelope

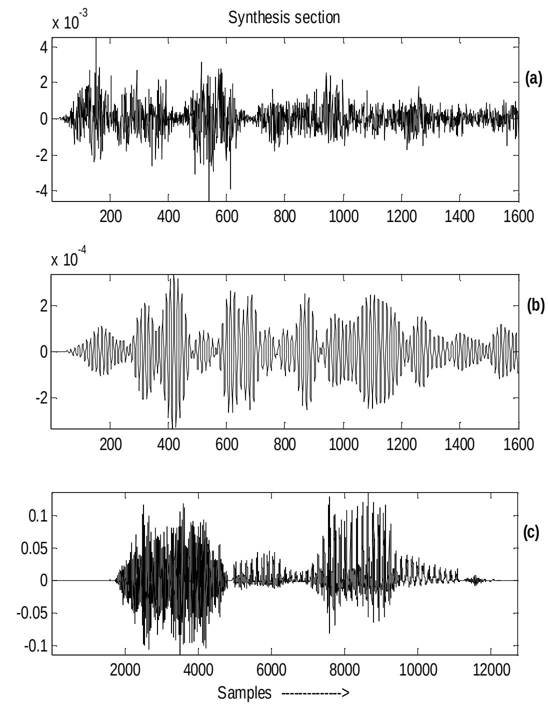

Fig 11.Uniform-bandwidth filters-based CI system - Synthesis section (a) Modulated envelope (unvoiced region) with random noise (b) corresponding filtered output and (c) synthetic speech for the word 'sugar'

filter bank-based CI model. For both word and sentence estimation channel variations on the speech signal are examined by varying the number of filters from 6-18 in steps of 3.

To validate the effectiveness of synthesis, spectral content of the input and the synthetic speech signal are compared using formant trajectories and to further verify closed-set listening test is conducted. Figs. 14 and 15 represent the input and synthetic speech signal for a sentence from TIMIT speech corpus respectively and their corresponding spectrograms with formant trajectories over riding on it. From the figures we observe that though the first two formants correlate to a greater extent to that of the input speech signal, further refinement, in terms of filter structure, is still required to modify the variations in the spectral content. To cross verify the effect of spectral variations a listening test is conducted as described in the following section.

4.4 Listening test

A closed-set listening test is conducted for all the synthesized vowels, generated using uniform-bandwidth filters-based CI model as well as for formant-filters based system. 17 normal hearing naive-listeners in the age group from 18 to 22 participated in the listening test. The perceptual quality of the synthetic speech is rated in 5-point grading varying from highly intelligible with a grade point of 5 to highly unintelligible with a rating of 1. It is observed that for the synthetic vowels generated using formant-filters outperformed the uniform-bandwidth based CI model in all aspects with a mean opinion score of 4.2 as opposed to 3.3 for the uniform-bandwidth CI model.

The perceptual quality of the synthetic words and sentences generated by uniform-bandwidth filterbank-based CI model is examined by conducting a listening test with the same set of listeners and the mean opinion score is found to be 3.3 for words and 3.1 for sentences. The reduction in quality of synthetic speech may be due to the variations in the formants with reference to the input signal as reflected in the spectral representations shown in Fig. 15. To further improve the perceptual quality of the speech signal an auditory-model-based CI system is developed as described in the following section.

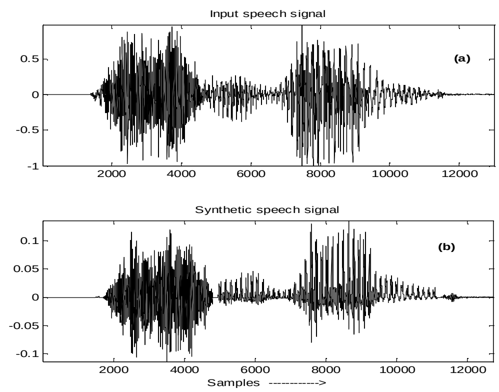

Fig 12.Input and synthetic speech signal estimated using uniform-bandwidth filter bank based CI-model for the word 'sugar' uttered by a male speaker.

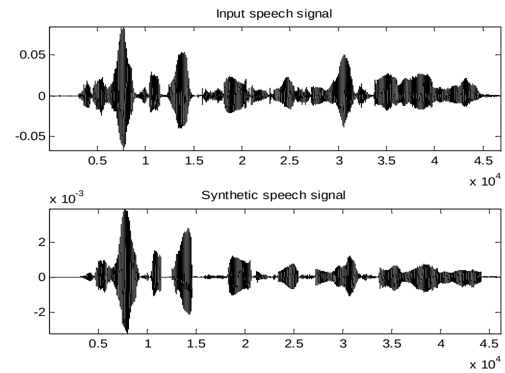

Fig.13 Input and synthetic speech signal for the sentence "she had your dark suit in greasy wash water all year" uttered by a female speaker from the TIMIT speech corpus, estimated using uniform bandwidth filter-bank based CI model.

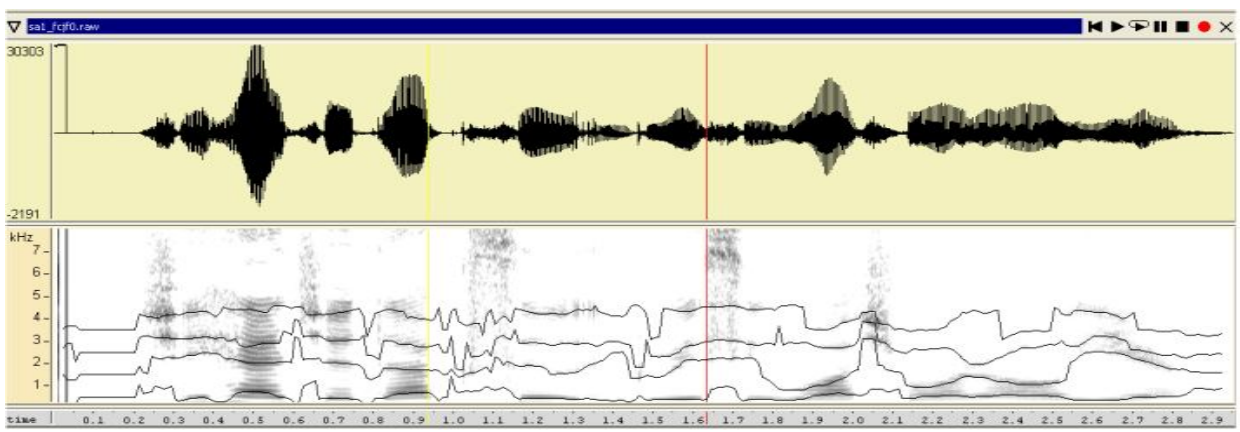

Fig. 14 Input speech signal and the corresponding spectrogram with first four formant trajectories shown as lines overriding on it.

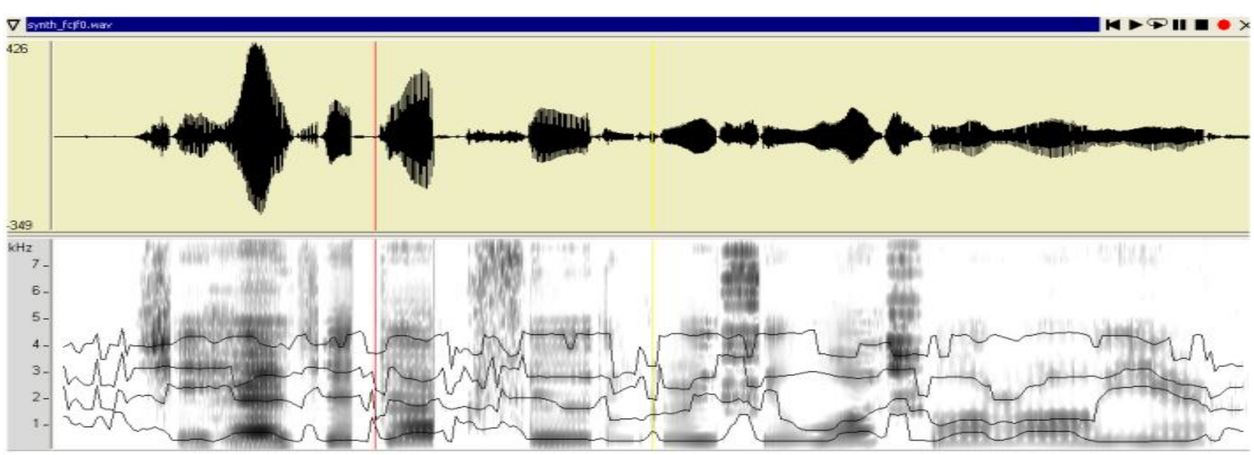

Fig. 15 Synthetic speech signal, estimated using uniform BW filters, and the corresponding spectrogram with first four formant trajectories overriding on it.

5 Auditory model-based acoustic CI simulations

In the auditory model based CI system bank of filters are designed based on the critical bands of the auditory system [20]. The bandwidth of the filters is computed based on equivalent rectangular bandwidth (ERB). If the center frequencies f_c of the filter are known then using the ERB the bandwidth of the filters can be computed as follows,

$$ERB = 24.7((0.0047f_c) + 1) \dots\dots\dots (4)$$

Table 1 lists the center frequencies and the corresponding ERB values of the bank of filters. The bank of filters are designed with center frequencies as shown in Table 1 and lower and upper cutoff frequencies of the filters are calculated as (*center frequency* ± ERB)} respectively.

5.1 Estimation of vowels, words and sentences

The analysis and synthesis sections of the auditory-model based CI systems are same as that of the uniform-bandwidth filters-based CI system as described in section 3 except for the center frequencies of the filters and the bandwidths considered. The speech data (vowels, words and sentences) collected from the speakers as described in section 2 are analyzed and synthesized using the

Table 1. Centre frequencies and corresponding ERBs

Center Frequency f_c *in Hz*	ERB in Hz	Center Frequency f_c *in Hz*	ERB in Hz
100	35.49	1480	184.45
300	57.08	1720	210.35
400	67.88	2000	240.58
510	79.75	2310	275.12
630	92.7	3150	364.7
770	107.81	3700	424.07
920	124	4400	499.63
1080	141.27	5300	596.77
1270	161.78	6400	715.51

CI model and the corresponding synthetic speech signals are generated. Figs. 16, 17, and 18 illustrate the input and synthetic speech signals for the vowel /a/, the word 'sugar' and a sentence from TIMIT speech corpus respectively.

5.2 Listening test

A closed-set listening test is conducted for all the synthesized vowels, generated using auditory-model based CI system. As described in section 4.4, the same set of 17 normal hearing naive-listeners participated in the listening test. The perceptual quality of the synthetic speech is rated in 5-point grading varying from highly intelligible with a grade point of 5 to highly unintelligible with a rating of 1. The mean opinion score is found to be 4.3 for words and 4.1 for sentences. As the critical-band based CI system is derived based on place theory [1] of the human auditory system the perceptual quality of the synthetic speech generated outperforms the speech intelligibility of the uniform bandwidth filters-based system as expected.

As described earlier to validate the results obtained from perceptual analysis, the spectral representations in terms of formant trajectories are plotted (refer to Figs. 19 and 20) for both input as well as for synthetic speech signal. From these Figs. it is observed that the higher formants (3rd and 4th, refer to Fig. 20) corresponding to the synthetic speech signal, apart from the lower formants, are also to a greater extent have trajectories closer to that of the input speech signal. This validates the improvement in perceptual quality of speech as observed in the intelligibility test.

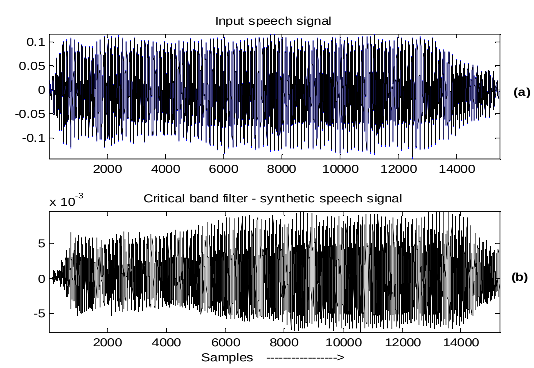

Fig. 16. Input and Synthetic speech signal for the vowel /a/ estimated by critical-band filters based CI system.

In the current work apart from waveform-based CI simulation models (uniform-bandwidth filter-based system, auditory-model based system), feature-extraction strategy based CI system is developed as described in the following section.

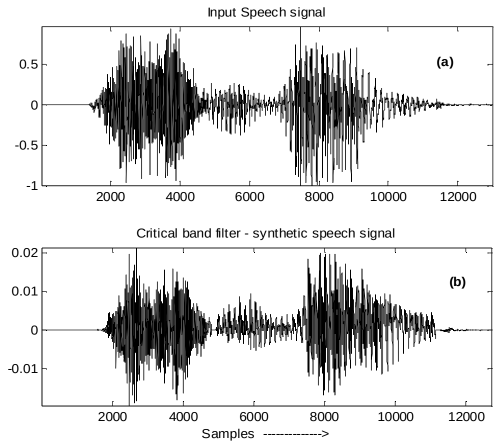

Fig. 17. Input and synthetic speech for the word 'sugar' estimated by critical-band filters-based CI model.

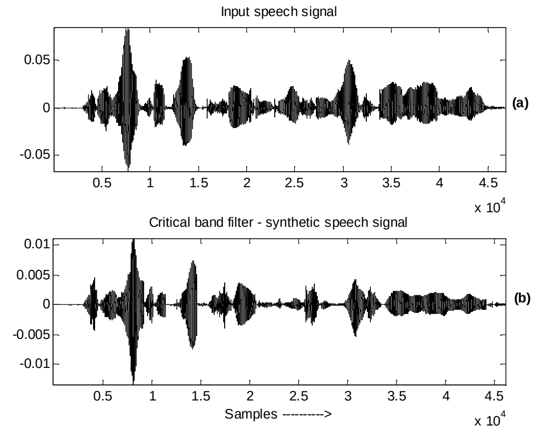

Fig 18 Input and Synthetic speech signal for the sentence "she had your dark suit in greasy wash water all year" from the TIMIT-speech corpus estimated using critical-band filters-based CI system.

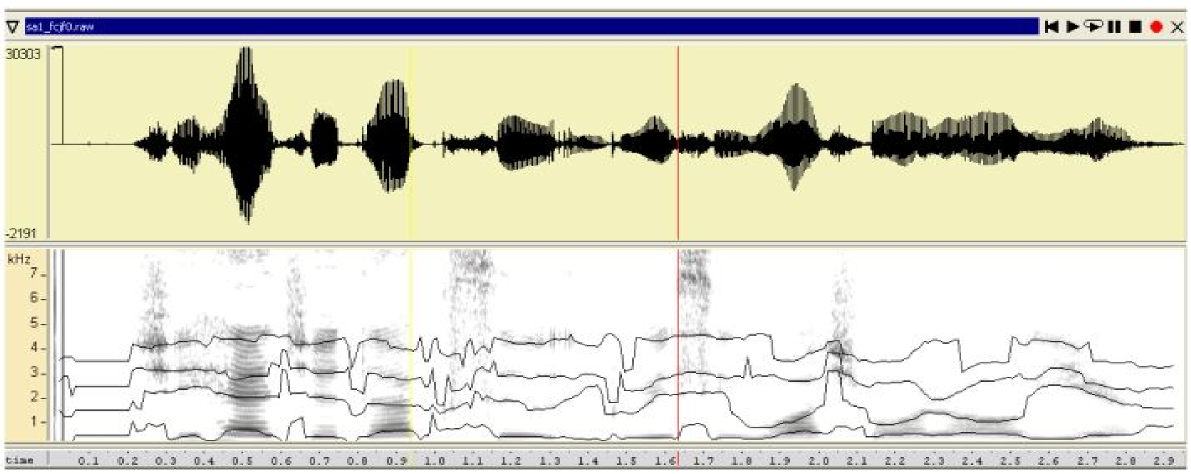

Fig. 19 Input speech signal and its spectrogram with first four formant trajectories overriding on it.

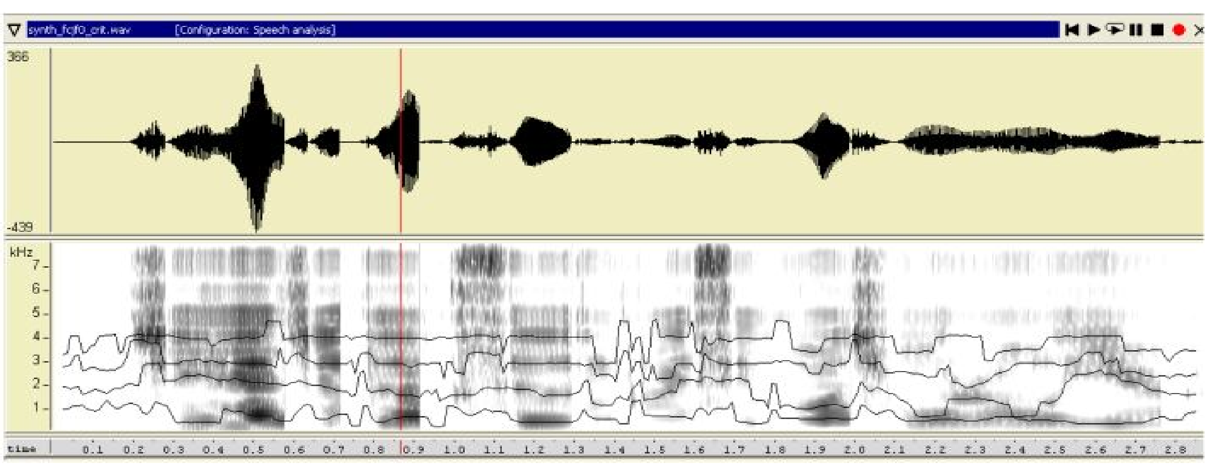

Fig. 20. Synthetic speech signal, estimated using critical-band filters-based CI system, and the corresponding spectrogram with first four formant trajectories overriding on it.

6 Mel-frequency cepstral coefficients-based analysis-synthesis system

In the feature extraction strategy apart from synthesizing vowels using formant synthesis technique as described in section 4.2, in the current work Mel-frequency cepstral coefficients (MFCCs) based analysis-synthesis system is developed. MFCCs mimic the human auditory system more closely than the linear filter-bank techniques.

The Mel-frequency cepstrum is a linear cosine transform of the log-power spectrum of the speech signal on a non-linear Mel-scale of frequency. The Mel-frequency cepstral coefficients are coefficients that collectively make up the Mel-frequency spectrum. In the analysis section of the system, the speech signal is analyzed in frames of 25ms each. After passing the signal through a pre-emphasis filter, the short-term Fourier transform is applied and its absolute value is squared to get the power-spectrum. The power spectrum is then frequency - warped to Mel-scale using triangular weighting functions. The conversion of linear frequency (*f*) to Mel-scale (*m*) is carried out by applying the following equation.

$$m = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \dots\dots\dots\dots (5)$$

For the current study 40 triangular filter functions are used, thus reducing the number of samples to 40 to represent each 25ms frame. This warping provides a better representation of the sound as the Mel-scale approximates the human auditory system. These spectral samples are then converted to cepstra by applying logarithm and discrete cosine transform of the warped samples followed by liftering to get 13 cepstral coefficients for each frame of data. The cosine transform also makes the coefficients linearly independent of one another. Fig. 21 illustrates steps involved in extracting Mel-frequency cepstral coefficients over a 25ms speech frame. The voiced, unvoiced and silence classification is performed based on GMM-tokenization procedure as described in Section 4.3.1. The MFCCs along with the voicing and pitch information is used to synthesize the speech signal.

In the synthesizer section, inverse liftering is performed to undo the liftering performed in the analysis section. The cepstral information is now reversed back to spectral information by applying the inverse discrete cosine transform. Exponential form of the resulting signal is computed. The power spectrum is then converted back to the normal frequency scale from the Mel-scale using the relation given by,

$$f = 700 * \left(10^{\frac{m}{2595}} - 1\right) \dots\dots\dots\dots (6)$$

Based on the voiced/unvoiced classification for each frame, a time domain signal of the length of the input speech is generated. This is achieved by modulating with the train of impulses separated by pitch period in case of voiced speech frames and random noise for unvoiced speech frames.

Using MFCC-based analysis-synthesis system speech data (vowels and LNT easy and hard words) collected from all the male and female speakers as described in Section 2 are analyzed and synthesized. 380 sentences uttered by 38 speakers (10 sentences each) from dialect-1 of TIMIT are synthesized using MFCC-based CI simulation system. Figs. 22 and 23 illustrate the input and synthetic speech signal for a word and sentence respectively. As described in the previous sections formant trajectories of input and synthetic speech signals are analyzed and is illustrated for a sentence from TIMIT speech corpus spoken by a female speaker in Figs. 24 and 25 respectively.

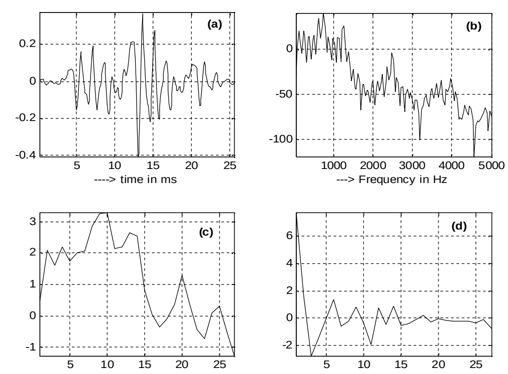

Fig. 21 Analysis section of MFCC - based system for a voiced speech frame. (a) Hamming windowed frame of speech, (b) Power spectrum of the speech segment, (c) Power spectrum plotted in Mel-scale (d) Mel-cepstrum

6.1 Listening test

As described in Section 5.2, 17 naive listeners participated in the closed-set listening test. The perceptual quality of the synthetic speech signal (words and sentences) generated by MFCC-based system are tested and compared with auditory-model based system. The mean opinion score for the synthetic words and sentences generated using MFCC-based system is 4.3 and 4.2 respectively. It is observed that as the filter structure of these methods

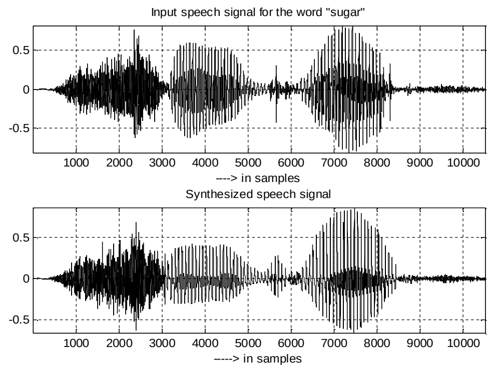

Fig. 22. Input and synthesized speech signal estimated using MFCC-based system for the word “sugar” uttered by a female speaker

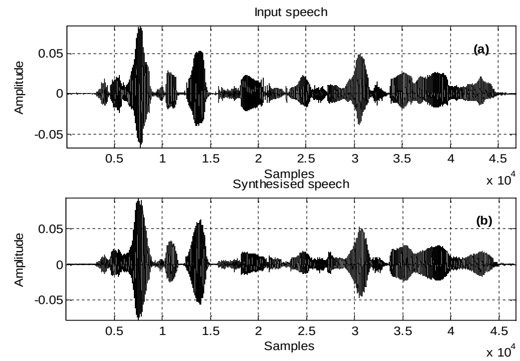

Fig. 23. Input and synthetic speech signal for the sentence "she had your dark suit in greasy wash water all year" from the TIMIT-speech corpus estimated by MFCC-based system.

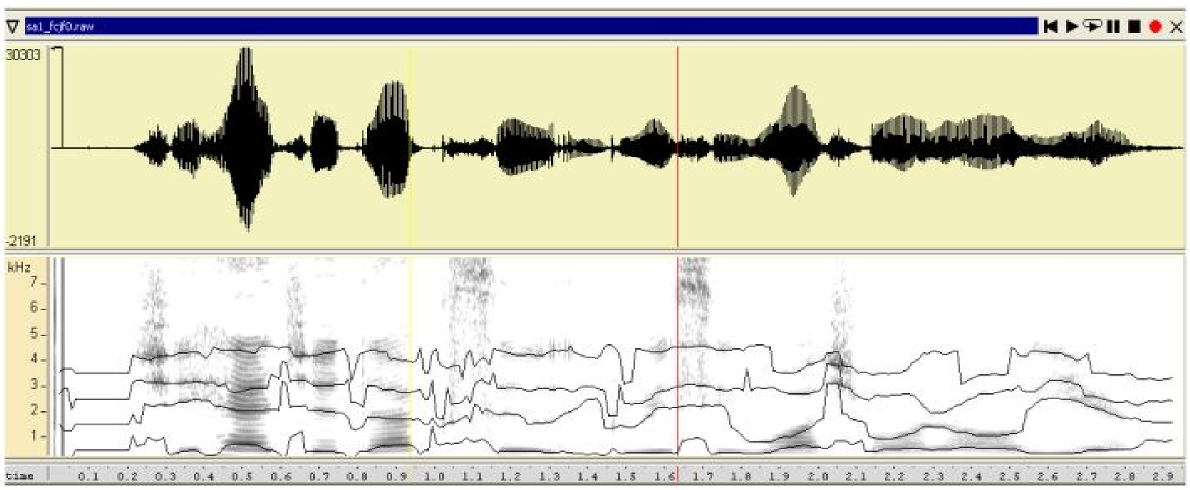

Fig. 24 Input speech signal and its spectrogram with first four formant trajectories overriding on it

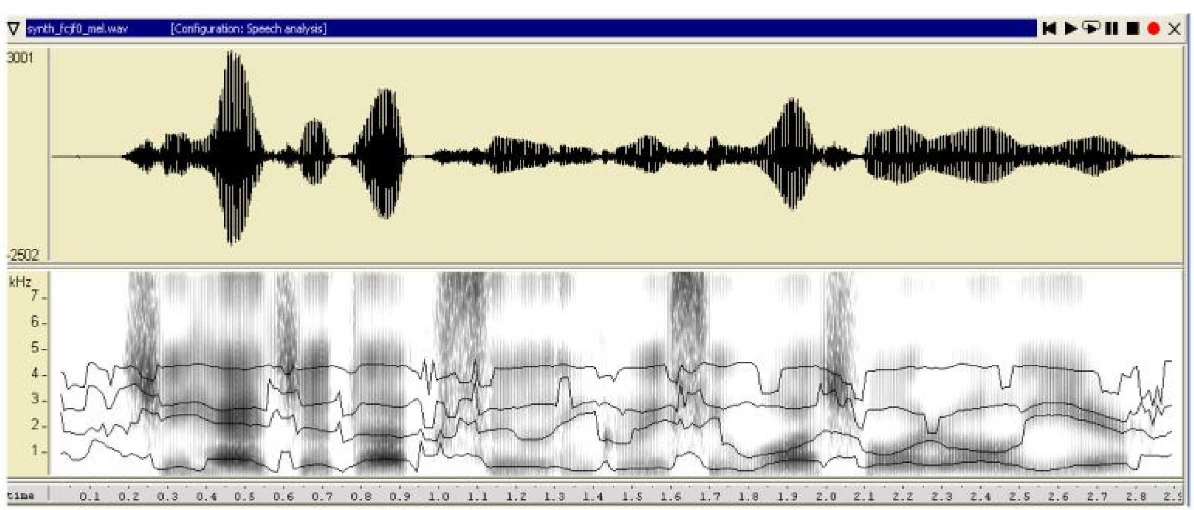

Fig. 25 Synthetic speech signal estimated using MFCC - based system and its spectrogram with first four formant trajectories overriding on it.

mimic the human ear to certain extent, the quality of the synthetic speech generated for both auditory-model based system and MFCC-based systems are found to be comparable and are reflected in the spectral content as seen in the formant trajectories.

7 Conclusions

In the present work cochlear implant models that are vocoder-centric are designed and evaluated. In particular, system specific parameters such as number of channels, type of filter banks, etc., are evaluated based on the perceptual quality of the synthetic speech generated in each method. The CI – models are developed for waveform as well as feature extraction strategies. The perceptual quality/ speech intelligibility of critical-band based system in the waveform strategy and the MFCC-based system in the feature-extraction strategy showed a comparatively better performance than the uniform-filter bank method. However, for hardware implementation waveform-coding technique would be a better choice as the receiver circuitry that is implanted will be less-complex for a waveform-based CI model compared to feature-extraction technique.

Acknowledgements: The authors thank all the listeners who participated in the listening test and the speakers from whom speech data are collected. Further authors would like to thank SSN trust for funding this project.

References:

[1] P.C. Loizou, ”Mimicking the human ear” *IEEE Signal Processing magazine***,** vol. 15, no. 5, Sep. 1998, pp. 101-130

[2] F. G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, “Cochlear Implants: System Design, Integration, and Evaluation” - *IEEE Reviews in Biomedical Engineering, v*ol. 1, 2008, pp 115-142.

[3] S. K. An, S. Park, S. B. Jun, C. J. Lee, K. M. Byun, J. H. Sung, B. S. Wilson, *S. J. Rebscher, S. H. Oh and* S. J. Kim, “Design for a Simplified Cochlear Implant System”, *IEEE Trans. on Biomedical Engineering*, Vol. 54, No.6 , Jun. 2007, pp. 973-982.

[4] H. J. McDermott, “How cochlear implants have expanded our understanding of speech perception” in *Proc. of the IEEE EMBC*, Vol. 20, No 5, 1998, pp. 2251-2256..

[5] J. Yao and Y. T. Zhang, “Bionic wavelet transform: a new time-frequency method based on an auditory model” - *IEEE Trans on Biomedical Engineering*, Vol. 48, No.8 Aug. 2001, pp. 856-863,

[6] P.C. Loizou, “Signal-processing techniques for Cochlear implants” *IEEE Engineering in Medicine and Biology*, May/Jun. 1999, pp. 34-46.

[7] J. Sit, A. M. Simonson, A. J. Oxenham, M. A. Faltys, and R. Sarpeshkar, “A Low-Power Asynchronous Interleaved Sampling Algorithm for Cochlear Implants That Encodes Envelope and Phase Information”, *IEEE Trans. on Biomedical Engineering*, Vol. 54, No. 1, Jan. 2007, pp. 138 - 149

[8] D. K. Eddington, W. M. Rabinowitz, and L. A. Dellzome, “Sound Processing for Cochlear Implants”, in *Proceedings of International IEEE EMBC*, 2001, pp. 3449-3452.

[9] C. Liu and Q. Fu, “Estimation of vowel recognition with cochlear implant simulations”, *IEEE Trans. on Biomedical Engineering*, Vol. 54, No.1, Jan. 2007, pp.74 – 81.

[10] K.I. Kirk, L.S. Eisenberg, A.S. Martinez AND M. Hay-McCutchen – “The lexical neighborhood test:- test-retest reliability and inter-list equivalency” – *Progress report no. 22 (Indiana university).*

[11] Fisher, W. M., G. R. Doddington and K. M. Goudie-Marshal, - “The DARPA speech recognition research database: Specifications and status, - *in Proc. DARPA workshop on Speech Recognition*, Feb. 1986, pp. 93 – 99.

[12] P. C. Loizou, “Speech processing in vocoder-centric cochlear implants” *Cochlear and Brainstem Implants. Adv Otorhinolaryngol. Basel, Karger*, vol 64, pp 109–143, 2006.

[13] B. Gold and N. Morgan, “Speech and audio signal processing - processing and perception of speech and music”. John Wiley and Sons. Inc., 2000.

[14] L. R. Rabiner and R. W. Schafer, Digital processing of speech signals. Prentice Hall, 1978.

[15] O'Shaughnessy, D., Speech Communications. IEEE Press, 2000.

[16] P. Vijayalakshmi, P. M. Kumar, Ra. V. Jayanthan and T. Nagarajan – “Cochlear implant models based on critical band filters” – *in Proceedings of IEEE TENCON 2009*, Singapore, Nov. 23 – 26, Nov. 2009, pp. 1 – 5.

[17] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation" *IEEE Trans. on Audio Electroacoust.*, pp. 367-377, Dec. 1972

[18] J. Makhoul, "Linear Prediction: A Tutorial Review" *Proc. of IEEE* Vol. 63, No.5, Apr. 1975, pp 561-580.

[19] A. Abraham, P. Vijayalakshmi and T. Nagarajan, "Pole-focused linear prediction-based spectrogram for coarticulation analysis"- in Proc. of IEEE Techsym 2010, pp. 94 – 97.

[20] Ra. V. Jayanthan, P. Vijayalakshmi and P. M.Kumar – "Auditory model based acoustic CI simulations for patients with profound hearing loss" – *in Proc. of Conference on Implantable Auditory Prosthesis (CIAP)* – 2009, pp. 228.