Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition

ASTIK BISWAS[1*], P.K.SAHU[1], ANIRBAN BHOWMICK[2], MAHESH CHANDRA [2]
[1] Dept. of Electrical Engineering, National Institute of Technology, Rourkela, India
[2] Dept. of ECE, Birla Institute of Technology, Mesra, Ranchi, India
astikbiswas@live.com[*], pksahu@nitrkl.ac.in, anirban.bhowmick@outlook.com
shrotriya@bitmesra.ac.in

Abstract: - In the presence of noise and sensor mismatch condition performance of a conventional automatic Hindi speech recognizer starts to degrade, while we human being are able to segregate, focus and recognize the target speech. In this paper, we have used auditory based feature extraction procedure Gammatone frequency cepstral coefficient (GFCC) for Hindi phoneme classification. To distinguish vowels from each other, we have analyzed frequency response curves of each vowel. Here we propose a new feature extraction technique by taking first three formant frequencies of each vowel along with their cepstral features to increase the phoneme classification performance in noisy condition. The classification performance achieved by the proposed features is compared with the standard MFCC and GFCC based features using a continuous density hidden Markov model (CDHMM) with a mixture of Gaussian distributions. To evaluate robustness of these features in noisy environment, the NOISEX database is used to add different types of noise into vowels in the range of 0 dB to 20 dB. Furthermore robustness of new set of feature has been evaluated in the sensor mismatch condition. The classification results show that under noisy background as well as the sensor mismatch condition the proposed technique achieves a better performance over standard cepstral based features.

Key-Words: - MFCC, GFCC, Formant, HMM, Phoneme Classification.

1 Introduction

The use of speech as a possible interface with computer/machines has gained popularity in the recent past. There are significant researches in developing robust speech recognition system in the past couple of decades. However, most of these systems developed by both in academics and in the industry are based on the Fourier transform for the analysis of speech signal. These systems have shown adequate recognition performance with clean data, keeping same acoustic conditions. Nevertheless, performance of Automatic Speech Recognition (ASR) still getting worse significantly in noisy environments and sensor mismatch conditions. Many algorithms have been proposed to address this problem, and they have achieved significant improvement in performance for stationary noise.

In Automatic Speech Recognition (ASR), front-end comprises the various feature extraction and noise compensation techniques, while different types of acoustic, language and pronunciation models are at back-end. Feature extraction is one of the important tasks for an ASR system. It is a process of extracting minimum information from a phoneme which gives maximum discrimination between phoneme classes. Thus, these features are important for overall recognition accuracy of an ASR system. Since speech recognition has to be performed into different environmental conditions, therefore, the features extracted must also be robust to background noise and sensor mismatch conditions. Although many feature extraction techniques have been proposed for speech recognition, some of the commonly used are Mel Frequency Cepstral Coefficients (MFCCs)[1], Linear Prediction based Cepstral Coefficients (LPCCs)[2], perceptual linear prediction, RASTA [3], [4], wavelet based feature [5], [6] etc.

There are numerous research have been carried out in different languages like English, Mandarin, Arabic etc. but little research have been carried out for Indian speech recognition[5], [7], [8] . Hindi is a mostly speaking Indian language belonging to the Indo–European family, which has retroflexion and gemination as important features [9]. Compared to the English language Hindi has more stop consonants and vowels [10]. To develop a reliable Hindi ASR these key characteristics of Hindi speech should be considered. Presently conventional Hindi ASR uses the standard feature extraction[11], [12] techniques, which are used in the English language which may not be optimum.

GFCC[13] have been proposed for the extraction of features for Mandarin and English speech. Specifically, GFCC is calculated by applying a cepstral analysis on the output of Gammatone filter bank, which was initially designed according to the frequency response of human cochlear filtering[14]. Furthermore, human listeners have the ability to focus and follow the particular speaker's voice in the multi-talker scenario or even in noisy conditions (like party, vehicle etc.) [15] as long as the signal-to-noise ratio (SNR) level is not going exceedingly low. The study of this perceptual process of humans is known as auditory scene analysis (ASA)[16]. Motivated by ASA studies, computational auditory scene analysis (CASA) looks to segregate and follow target speech from a complex auditory scene[17].

Formant frequencies [18]–[20]have been used in speech recognition by many researchers in the past. Voiced sounds (vowels, for example) produced by acoustically filtered quasi-periodic air pulses as they propagate along the vocal tract. The resonant frequencies of a vocal tract are exactly the formant location while pronouncing a voiced sound. The first three formants are sufficient to distinguish vowel form consonants. These formant frequencies are helpful for phoneme classification when acoustic conditions are not same or in the presence of noise. In the succeeding sections, we propose a new feature extraction method by taking formant frequencies of each vowel along with their cepstral features. Further, these features are likely to be more robust to sensor mismatch condition and in the presence of noise as compared to standard feature extraction techniques. Some of researcher also have mentioned the use of software agent in artificial intelligence [21]–[23] but here we have limited our focus with HMM.

This paper is organized as follows: section 2 and 3 gives an overview auditory based feature and formant analysis on Hindi vowel. Section 4 presents the preparation of Hindi speech database. Finally, section 5 and section 6 gives the results and conclusion of the experiments performed.

2 Auditory Based Feature

In CASA system, a standard model for time-frequency (T-F) analysis involves a series of Gammatone filters[17], which decomposes an input signal into the T-F domain. In computational auditory models, the peripheral filtering in the cochlea is typically described by Gammatone filters. Psychophysical observations of the auditory perceptual study are used to design Gammatone filters, which follows frequency response of human cochlea [14]. The impulse response of a Gammatone filter centered at frequency f is given by: -

$$g(f,t)=\begin{cases} t^{a-1}e^{-2\pi bt}\cos(2\pi ft), & t\geq 0. \\ 0, & else \end{cases} \quad (1)$$

where, t denotes time; a is the filter order is equal to 4; rectangular bandwidth b increases with the center frequency f. A bank of 64 filters have been used in this experiment whose center frequency f ranging from 50 Hz to 8000 Hz. These center frequencies of Gammatone filter band are correspondingly shared along the Equivalent Rectangular Bandwidth (ERB) scale and the filters with higher center frequencies react to wider frequency ranges.

The GFCC implementation proposed in [11] have used a series of 128 filters having center frequency f ranges from 50 Hz to 8000 Hz and then extracted frames from the GF outputs by down-sampling y (t; f_m) to 100 Hz where f_m is the central frequency of the m^{th} GF. This approach results in high variation even with a low-pass filter. We have used an average approach [13] which uses a window covering K points and shifting every L points to frame $y\ (t;\ f_m)$. For the n^{th} frame, the average value of y (t; f_m) within the window $t\in\left[nL,nL+k\right]$ is computed as the m^{th} component:

$$\overline{y}(n;m)=\frac{1}{K}\sum_{i=0}^{K-1}\gamma(f_m)\left|y(nL+i;\,f_m)\right| \quad (2)$$

where $\gamma(f_m)$ is a center frequency-dependent factor, and m is the index of the channel whose central frequency is f_m. Here we took only the magnitude of complex numbers. The length of Hindi vowel generally lies between 28 ms-130 ms. Hence, we have chosen K = 256, L = 160, and a bank of 64 filters has been used (i.e.$0\leq$ m< 64). This means each frame corresponds to a 64-dimension vector $\overline{Y}(n)=[\overline{y}(n;0),\overline{y}(n;1),.....,\overline{y}(n;63)]^T$. For 16 kHz signals, these settings result in 100 frames per second, exactly the same as the down-sampling approach and the usual frame rate of MFCC. The resulting matrix $\overline{Y}(n,m)$ provides a TF representation of the original signal, and is often referred as a Cochleagram [13]. A typical Cochleagram of Hindi vowel */i/* is shown in Figure 1.

A time frame of these cochleagram is known as a Gammatone feature (GF) and comprises of 64 frequency components. Because of overlapping among neighboring filter channels, the GF are

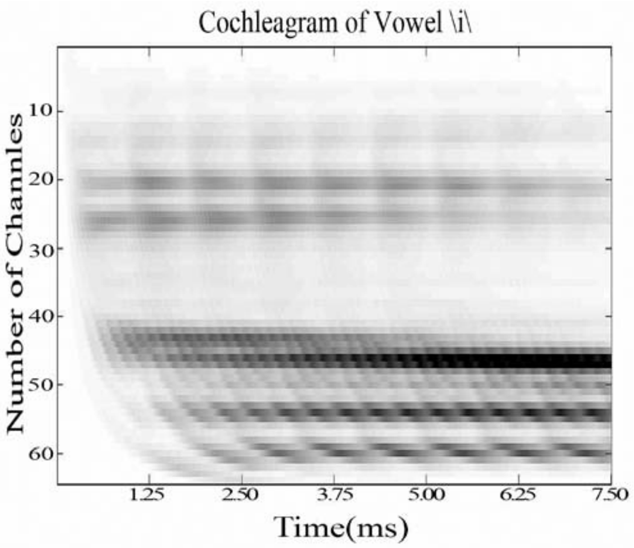


Figure 1. Cochleagram of Hindi vowel \i\

highly correlated with each other. Thus to de-correlate this GF components we have applied a discrete cosine transform (DCT) as mentioned in [13]. The resulting coefficients after applying DCT are known as GFCC[13]. Note that before applying DCT, we have performed logarithm on GF as usually adopted in the MFCC processing. This logarithm operation leads to more suitability in numerical processing. The following equation presents the exact cepstral form:

$$F(n,u)=\left(\sqrt{\frac{2}{M}}\right)\sum_{i=1}^{M}\{\frac{1}{3}\log(\bar{y}(n,i))\cos[\frac{\pi u}{2N}(2i-1)]\} \quad (3)$$

where the total number of channels is given by M is equal to 64, and u range from 0 to 63 accordingly. Hence finally 64 GFCC features have been derived for each frame whose dimension is much larger than standard feature vectors used in a typical ASR. In the previous study by [13], the 30 lowest order GFCC coefficients were used as a feature vector. In our case, we have found that first 27 lower coefficients captured almost all the information while the higher order GFCCs above the 27th are close to zero. Hence, we have used the 27-dimensional GFCC as an acoustic feature vector. Furthermore, in order to keep 13 features (analogous to the number of MFCC features) Linear discriminant analysis (LDA) [24] has been applied for dimensionality reduction to get the final 13-GFCC feature vector for each frame.

Along with the static feature, a dynamic information from speech signal is also captured. First-order delta coefficients have been calculated to incorporate temporal information[11]. Delta features ΔF at time frame n is given as,

$$\Delta F(n,u)=\frac{\sum_{k=1}^{K}k(F(n+k,u)-F(n-k,u))}{2\sum_{k=1}^{K}k^2} \quad (4)$$

Here, F refers to static GFCCs; neighboring window index refers k; K refers the half-window length, and it is set to 2.

3 Formant Analysis

A well-accepted model of speech production characterizes the vocal tract as a tube or concatenation of tubes of varying cross-sectional areas. In between the excitation and the output at the resonant frequencies most of the energy is transferred. The formant frequencies are the resonant frequencies of the speech signal. They correspond to high-energy regions of the spectrum and can be identified by the peaks of the spectral envelope.

The estimation of formants on a frame-by-frame basis is achieved by using LPC, and then finding the peaks on this envelope. The first N peaks are assigned to the first N formants. There are typically about three resonances of significance below 3500 Hz. The first formant, F1, is the lowest resonant frequency. The lowest two or three formants are usually sufficient to identify specific phonemes, while the location of the higher formant is generally contains the speaker dependent information. The problem arises while estimating the formant for low-level voiced sounds, and the difficulty of defining the formant for unvoiced or silence regions. The main difference between vowels and consonants is that vowels resonate in the throat, while consonants are produced by restricting air flow over the articulators. Consonants also resonate in the nasal passage (to a small extent vowels do too, but can be ignored in simple models).

Today, virtually all high-performance speech recognition systems are based on filterbank analysis. Nevertheless, the performance of high- performance ASR degrades in presence of high value of noise or mismatch between training and testing conditions. There might be specific aspects due to which formant-based parameters are attractive, as listed below [25].

- Formants are considered to be robust against channel distortions and noise.
- Formant parameters might useful to overcome the problem of a mismatch between training and testing conditions.

- There is a close relation of formant parameters to model based approaches to speech perception and production.

3.1 Current Research in Vowel Recognition

The use formant frequencies to recognize the Arabic vowels is described in [26]and [27]. In [26] authors have described the segmentation and identification of Arabic vowels in continuous speech using formant frequencies. They have worked with 1000 Arabic vowels and achieved 90% recognition efficiency. Alotaibi et al. [27] have researched with Arabic vowels by using their TF domain characteristics and formant frequencies. They have demonstrated an ASR based on Hidden Markov Model (HMM) with recognition efficiency was about 91.6%.

In recent studies of English vowels by Kodandaramaiah et al. [28] have demonstrated the effectiveness of formants to classify the vowels based. They have achieved with 80% to 95% of accuracy based on Euclidean distance. Kocharov [29] has developed an ASR to recognize the Russian vowels which is based on synchronization with the pitch period. They have got the recognition efficiency of 87.70% for isolated vowel recognition task and 83.93% for the vowels within a word.

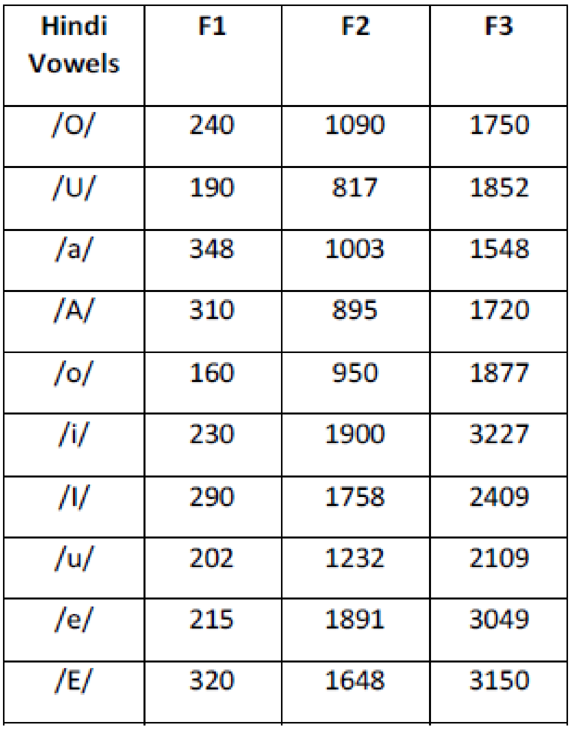
3.2 Formants Analysis on Hindi Vowel

We have analysed the formant frequencies of first three formants of Hindi vowels simultaneously using 11th order Linear Prediction Analysis (LPC) [18]. The positions for the first three formants of a vowel aren't random. The average locations of first three formants of typical 10 Hindi vowels given in table 1. For an open vowel (such as */a/*), the first formant F1 has a higher frequency while close vowel (such as */i/* or */u/)* has a lower frequency and the second formant F2 has a higher frequency for a front vowel (such as */i/*) and a lower frequency for a back vowel (such as */u/*).

Next we have examined the effect of noise on these formant locations of Hindi Vowels. The TIFR- CEERI Hindi continuous speech database is used to study the effect of noise on formants frequency. We have taken five types of noise form NOISEX database. The speech signal is degraded by five different types of noise with an average SNR in the range from 0 to 20 dB. Robust formant tracking [30] method has been used to locate formant locations in speech signal. The procedure contains pre-cleaning of the spectral amplitude of speech with formant estimation followed by smoothing of the formant trajectories with Kalman filters. Figure 2, 3 and 4 illustrates the effect of different noise with 0 dB SNR on the formant locations respectively. The response curves shown below was obtained from 10 utterances by an adult male native Hindi speaker in TIFR-CEERI Hindi database.

Note that, in the presence of high noise also, first three formant locations do not alter too much. Therefore, these formant locations might be useful for speech recognition in the presence of noise or when acoustic conditions are not same, and this influences us to use the formant for Hindi vowel classification.

Table 1. Formant Frequencies for typical Hindi vowel



Hindi Vowels	F1	F2	F3
/O/	240	1090	1750
/U/	190	817	1852
/a/	348	1003	1548
/A/	310	895	1720
/o/	160	950	1877
/i/	230	1900	3227
/I/	290	1758	2409
/u/	202	1232	2109
/e/	215	1891	3049
/E/	320	1648	3150

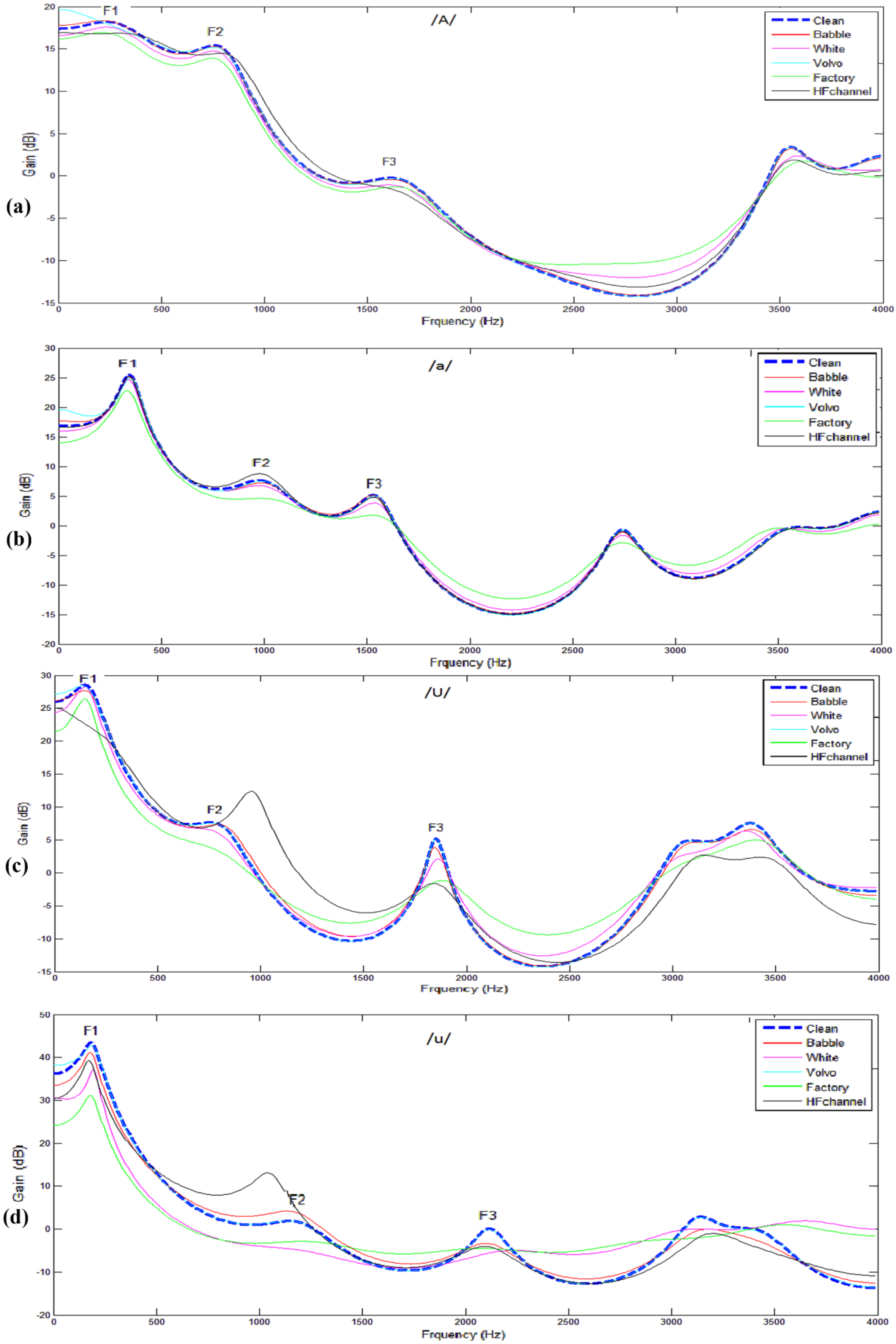


Figure 2. Effect of noise on the formant frequencies. (a) vowel /A/, (b) vowel /a/, (c) vowel /U/,

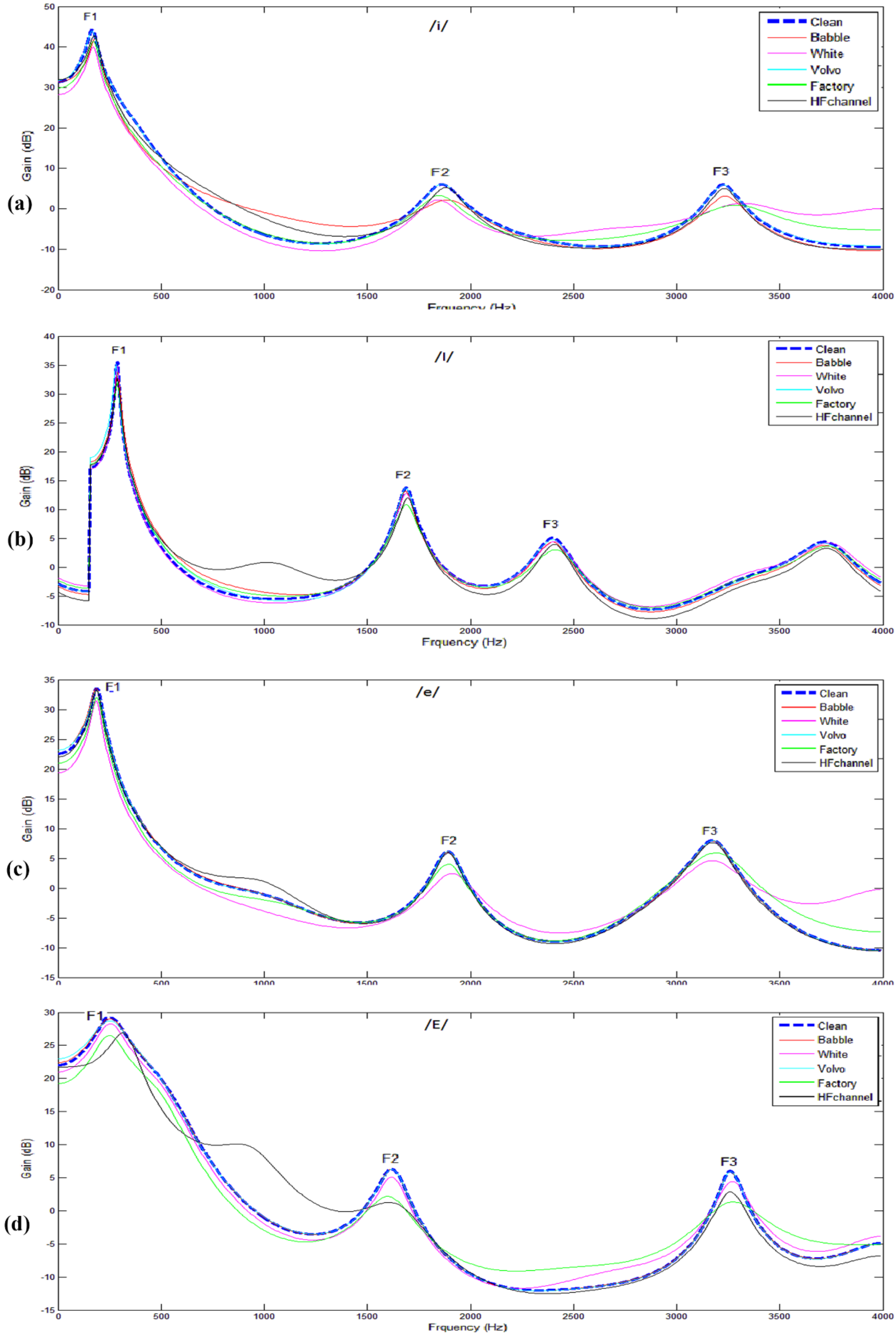


Figure 3. Effect of noise on the formant frequencies. (a) vowel /i/, (b) vowel /I/, (c) vowel /e/,(d) vowel /E/

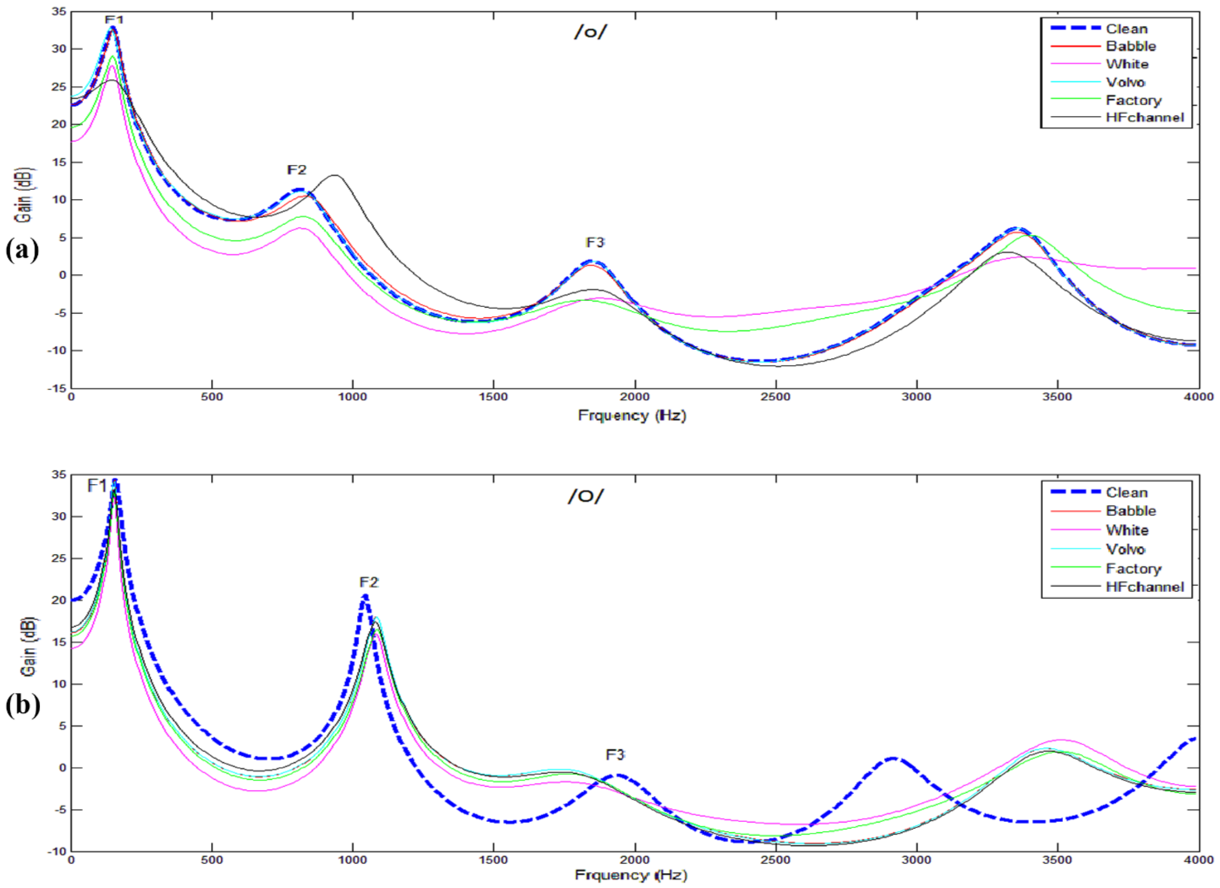


Figure 4. Effect of noise on the formant frequencies. (a) vowel /A/, (b) vowel /a/, (c) vowel /U/, (d) vowel /u/

4 Hindi Speech Database

A Hindi speech database [31] was used to extract the phonemes for classification. From this database, a total of 90 speakers was selected out of which 58 were male and 32 females. Each speaker had 10 phonetically rich sentence utterances, out of which two sentences were common for all the speakers. The phonetically rich sentences were designed at TIFR, Mumbai, India. The speech corpus was recorded at CEERI, New Delhi, India with 16 kHz sampling frequency and stored in the 16-bit PCM-encoded waveform format in monomode using two microphones: one high quality close-talking directional microphone and another desk-mounted at a distance of 1 meter omni-directional microphone. Phoneme boundaries are provided in the database from the spoken sentences which were manually segmented. Here we have worked with 10 typical Hindi Vowels. Vowels were extracted using these labels provided in the database. For our classification purpose, Vowels were grouped based upon the place of articulation (the list is shown in Table 2).

Noisy phonemes were generated by adding different levels of noise from the NOISEX-92 database. Suitable scaling was applied to obtain the desired signal-to-noise ratio (SNR) without clipping the resultant audio. Three different types of noise (Babble, Factory and White) were selected for robustness evaluation of the proposed features.

Table 2. The set of Hindi vowels showing their place of articulation

Front Vowel	Mid Vowel	Back Vowel
/I/ [ई]	/a/ [अ]	/u/ [उ]
/i/ [इ]	-	/U/ [ऊ]
/e/ [ए]	-	/o/ [ओ]
/E/ (/ai/) [ऐ]	-	/O/ [औ]
-	-	/A/ [आ]

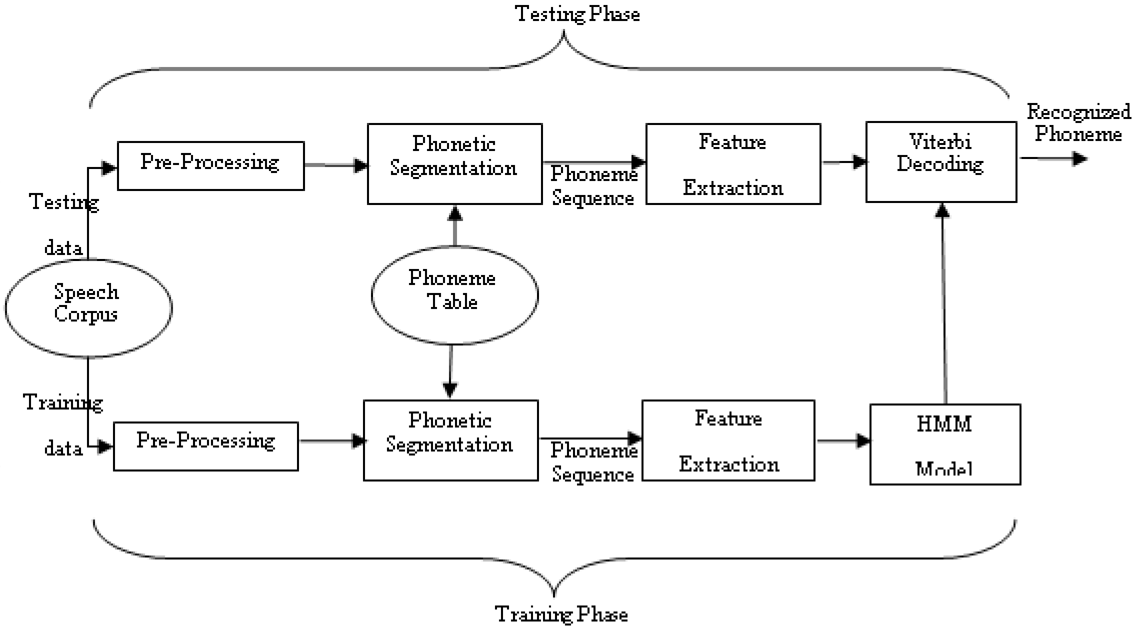


Figure 5. Experimental setup of HMM based phoneme recognizer

5 Experimental Setup and Results

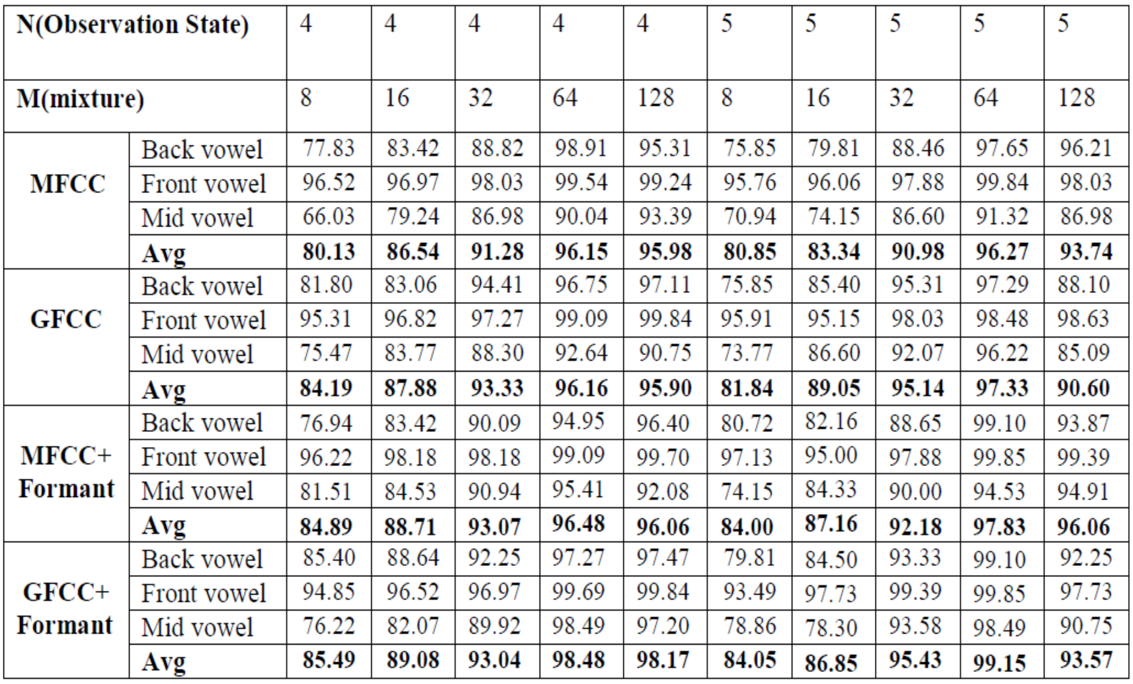
The analog speech signal was digitized with the sampling rate of the order of 16 kHz and using 16 bits/sample for quantization. The signal was pre-emphasized using the 0.97 value for its coefficient, to ensure that all formants of acoustic signals have similar amplitudes so that they get equal importance in subsequent processing stages. Finally, 26 acoustic features, including delta features were derived on frame by frame basis (using 16 ms window and 10 ms overlapping) by applying filter bank approach like MFCC and GFCC. We have used a filter bank of 27 overlapped triangular filters to calculate the static MFCCs. The procedure of finding GFCC features is described earlier in section 2. To get formant based feature, first three formants frequency is concatenated with cepstral based features. Finally, 29 acoustic features were derived to get formant based feature like MFCC+Frmnt and GFCC+Frmnt. Figure 5 shows the experimental setup used in this research.

To study our ASR performance, we have divided our work in the following sub-categories:-

- *Speaker dependent (case 1)*: All the 90 speakers have been used for training, while 20 speakers (13 males, 7 females) have been used for testing. All speech files used for training and testing both were recorded with close talk microphone.
- *Speaker independent (case 2)*: 70 speakers have been used for training, while rest 20 speakers (13 males, 7 females) have been used for testing. All speech files used for training and testing both were recorded with close talk microphone.
- *Speaker dependent sensor mismatch condition (case 3):* All the 90 speakers have been used for training, while 20 speakers (13 males, 7 females) have been used for testing. Speech utterances have been used for training were recorded with close-talk microphone, While testing speech utterances were recorded with desk-mounted microphone.
- *Speaker independent with sensor mismatch condition (case 4):* 70 speakers have been used for training, while 20 speakers (13 males, 7 females) have been used for testing. Speech utterances have been used for training were recorded with close-talk microphone, while testing speech utterances were recorded with desk-mounted microphone.
- *Speaker independent with noisy Environment (case 5):* 70 speakers have been used for training, while 20 speakers (13 males, 7 females) have been used for testing. All speech files used for training and testing both were recorded with close talk microphone. Babble, White and factory noises from NOISEX database have been used to get noisy speech.

Different HMM [18], [32], [33] models were experimentally observed in order to find the best model for Hindi vowel classification task. For each HMM model, the classification results were given as a function of percentage classification of three vowel classes (i.e. front vowel, mid vowel and back vowel). For each vowel class, the state number has been fixed to four and five; and 8, 16, 32, 64 and 128 Gaussian mixtures with a full covariance matrix were used to get the best acoustic HMM model. The Viterbi decoding algorithm was used for training by taking into account the practical implementation issues, such as scaling, multiple observation sequences and initial parameter estimates, which are explained by Rabiner et al[18]. During the training, training terminated automatically when the highest accuracy is obtained for the validation data recognition. This termination also prevents the over training of HMMs. Again, the Viterbi algorithm was applied in the testing phase to classify the phonemes into their respective classes.

Table 3. Classification results of different vowel class with clean speech (speaker dependent)



N(Observation State)		4	4	4	4	4	5	5	5	5	5
M(mixture)		8	16	32	64	128	8	16	32	64	128
MFCC	Back vowel	77.83	83.42	88.82	98.91	95.31	75.85	79.81	88.46	97.65	96.21
	Front vowel	96.52	96.97	98.03	99.54	99.24	95.76	96.06	97.88	99.84	98.03
	Mid vowel	66.03	79.24	86.98	90.04	93.39	70.94	74.15	86.60	91.32	86.98
	Avg	**80.13**	**86.54**	**91.28**	**96.15**	**95.98**	**80.85**	**83.34**	**90.98**	**96.27**	**93.74**
GFCC	Back vowel	81.80	83.06	94.41	96.75	97.11	75.85	85.40	95.31	97.29	88.10
	Front vowel	95.31	96.82	97.27	99.09	99.84	95.91	95.15	98.03	98.48	98.63
	Mid vowel	75.47	83.77	88.30	92.64	90.75	73.77	86.60	92.07	96.22	85.09
	Avg	**84.19**	**87.88**	**93.33**	**96.16**	**95.90**	**81.84**	**89.05**	**95.14**	**97.33**	**90.60**
MFCC+ Formant	Back vowel	76.94	83.42	90.09	94.95	96.40	80.72	82.16	88.65	99.10	93.87
	Front vowel	96.22	98.18	98.18	99.09	99.70	97.13	95.00	97.88	99.85	99.39
	Mid vowel	81.51	84.53	90.94	95.41	92.08	74.15	84.33	90.00	94.53	94.91
	Avg	**84.89**	**88.71**	**93.07**	**96.48**	**96.06**	**84.00**	**87.16**	**92.18**	**97.83**	**96.06**
GFCC+ Formant	Back vowel	85.40	88.64	92.25	97.27	97.47	79.81	84.50	93.33	99.10	92.25
	Front vowel	94.85	96.52	96.97	99.69	99.84	93.49	97.73	99.39	99.85	97.73
	Mid vowel	76.22	82.07	89.92	98.49	97.20	78.86	78.30	93.58	98.49	90.75
	Avg	**85.49**	**89.08**	**93.04**	**98.48**	**98.17**	**84.05**	**86.85**	**95.43**	**99.15**	**93.57**

Table 3 shows the classification results for the different HMMs for case 1 and Table 4 shows the results for case 2. As expected, speaker dependent results are better than the speaker independent results. From the results, it can be notable that the performances of the HMMs getting better with number of Gaussian mixture components increases at the cost of more computational complexity. Among the all ten different phoneme models, HMMs with five states and 64 Gaussian mixtures are found to be the best one for both speaker dependent and independent cases. Note that, performance of HMM with more than 64 mixture components degrades because of insufficient training data, and computational complexity also increases. From table 3 and table 4, it can be seen that with clean data, most of the time auditory based feature GFCC outperforms MFCC, while formant based proposed feature have shown better classification efficiency than cepstral based features. GFCC with formant based feature has shown near about 99% classification efficiency in the both cases.

To evaluate the robustness of the extracted features in sensor mismatch condition, speech files recorded with desk mounted omnidirectional microphone have been used. Table 5 and table 6 shows the classification results for the different HMMs mentioned in case 3 and case 4 respectively. Here also, HMMs with 5 states and 64 Gaussian mixtures are the best one for both speaker dependent and independent cases. Here also average classification efficiency of speaker dependent case are marginally better than the speaker independent case. Table 5 and table 6 shows that GFCC has shown better classification result compared to MFCC in most of the time. The performance is further improved with formant based cepstral features like MFCC+Formant frequencies and GFCC+Formant frequencies. In the sensor mismatch condition, the average classification result of the desk-mounted microphone considerably drops to less than 80% compared to close talk

Table 4. Classification results of different vowel class with clean speech (speaker independent)

N(Observation State)		4	4	4	4	4	5	5	5	5	5
M(mixture)		8	16	32	64	128	8	16	32	64	128
MFCC	Back vowel	75.25	82.24	86.90	96.50	93.60	74.60	79.81	86.84	93.09	82.25
	Front vowel	95.24	94.95	96.07	98.45	97.98	94.50	96.06	97.15	98.17	96.95
	Mid vowel	65.21	77.52	84.85	88.20	92.60	68.35	74.15	86.98	93.22	82.95
	Avg	**78.57**	**84.90**	**89.25**	**94.38**	**94.73**	**79.15**	**83.34**	**90.32**	**94.83**	**87.38**
GFCC	Back vowel	81.00	82.35	93.50	95.40	96.15	74.20	83.85	94.5	96.00	87.10
	Front vowel	94.65	95.88	97.27	98.62	98.90	94.78	93.90	96.75	97.10	97.08
	Mid vowel	74.55	82.40	87.25	91.73	89.31	72.75	85.05	90.89	94.70	83.90
	Avg	**83.40**	**86.88**	**92.67**	**95.23**	**94.78**	**80.58**	**87.58**	**94.05**	**95.93**	**89.33**
MFCC+ Formant	Back vowel	82.25	89.60	94.00	94.90	79.14	81.44	87.46	98.35	96.25	95.60
	Front vowel	97.05	97.15	98.15	98.35	96.27	93.88	96.92	98.20	98.90	97.81
	Mid vowel	82.95	89.55	94.92	90.65	73.50	83.42	89.14	92.89	90.45	86.00
	Avg	**87.42**	**92.10**	**95.69**	**94.63**	**82.97**	**86.25**	**91.17**	**96.48**	**95.20**	**93.13**
GFCC+ Formant	Back vowel	84.05	86.95	91.07	96.12	96.42	78.42	83.37	92.42	98.62	91.15
	Front vowel	94.10	95.32	96.00	98.65	98.30	92.64	96.75	98.70	99.00	96.95
	Mid vowel	73.94	81.00	88.24	97.50	96.30	77.72	78.00	92.92	97.88	88.86
	Avg	**84.03**	**87.76**	**91.77**	**97.42**	**97.01**	**82.93**	**86.04**	**94.68**	**98.50**	**92.32**

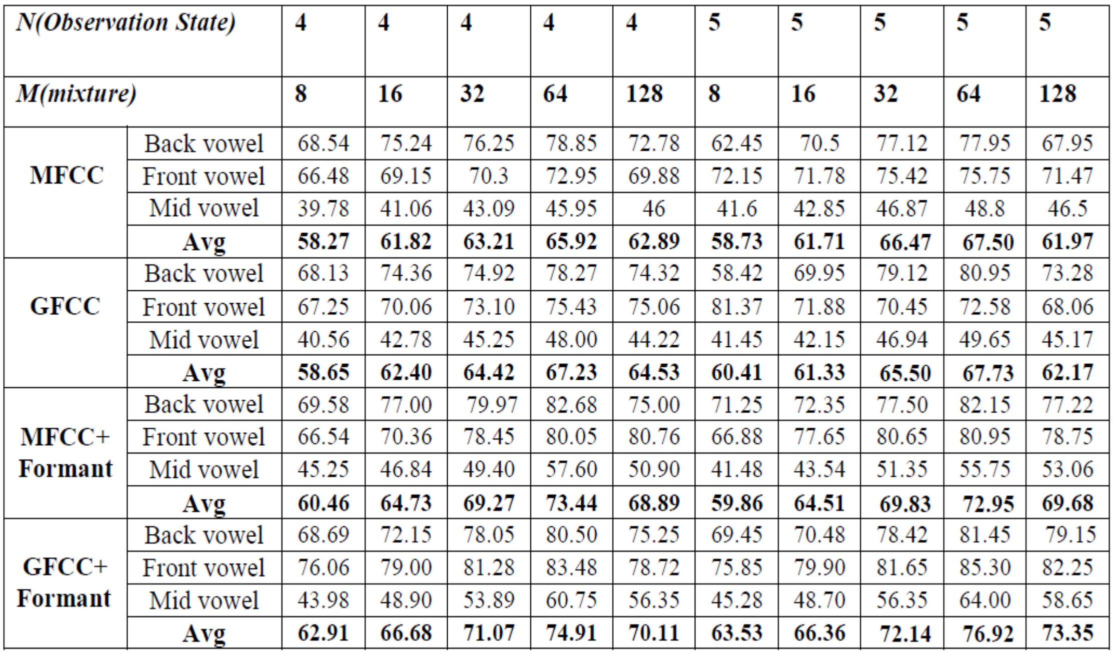
microphone. This is because of, the omnidirectional microphone is sensitive to sound rather than the direction of the incoming sound. Thus, it picks up the wanted sound produced by the speaker as well as unwanted background noise.

Table 5. Classification results of different vowel class with desk mounted microphone (speaker dependent)

N(Observation State)		4	4	4	4	4	5	5	5	5	5
M(mixture)		8	16	32	64	128	8	16	32	64	128
MFCC	Back vowel	70.23	76.39	77.25	80.78	74.62	63.80	71.24	78.65	80.91	68.62
	Front vowel	68.65	70.20	71.35	74.00	71.22	73.50	73.78	76.90	76.78	72.35
	Mid vowel	40.72	42.60	44.89	47.24	47.28	42.83	44.40	48.25	49.60	47.41
	Avg	**59.87**	**63.06**	**64.50**	**67.34**	**64.37**	**60.04**	**63.14**	**67.93**	**69.10**	**62.79**
GFCC	Back vowel	65.22	75.50	76.24	79.94	75.98	59.82	71.24	80.05	82.71	74.60
	Front vowel	67.77	71.12	74.00	76.78	75.81	82.45	72.25	71.60	73.82	69.15
	Mid vowel	41.53	43.56	46.81	49.30	45.51	42.83	43.56	48.25	51.00	46.41
	Avg	**58.17**	**63.39**	**65.68**	**68.67**	**65.77**	**61.70**	**62.35**	**66.63**	**69.18**	**63.39**
MFCC+ Formant	Back vowel	70.22	78.56	81.80	84.61	76.22	72.12	73.51	78.74	83.60	78.90
	Front vowel	67.77	72.47	79.60	81.21	82.75	68.95	79.43	81.54	82.31	80.03
	Mid vowel	46.52	47.68	50.36	59.34	52.98	42.31	44.60	52.35	57.24	54.90
	Avg	**61.50**	**66.24**	**70.59**	**75.05**	**70.65**	**61.13**	**65.85**	**70.88**	**74.38**	**71.28**
GFCC+ Formant	Back vowel	70.45	73.6	79.45	82.7	77.09	70.27	71.53	80.54	83.58	80.72
	Front vowel	77.15	80.15	82.9	84.56	80.78	77.76	81.24	83.05	86.98	83.35
	Mid vowel	45.47	50.5	55.32	62.45	57.51	45.28	49.25	58.85	65.45	58.65
	Avg	**64.36**	**68.08**	**72.56**	**76.57**	**71.79**	**64.44**	**67.34**	**74.15**	**78.67**	**74.24**

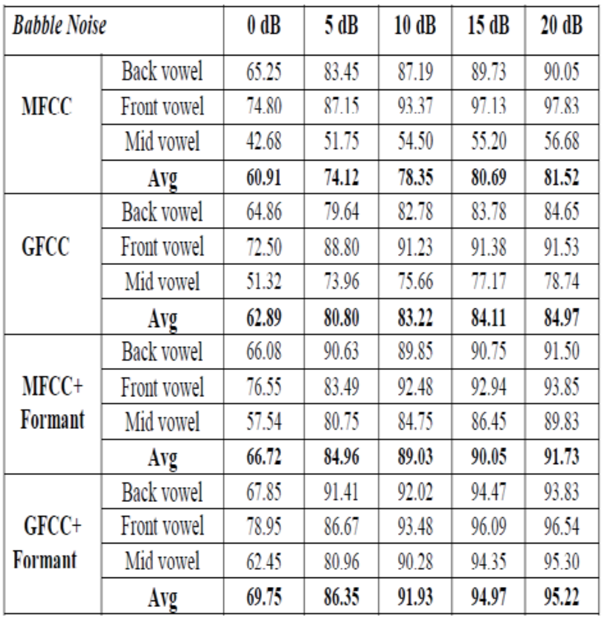
Next, we have evaluated the performance of the proposed feature in the noisy environment as mentioned in case 5. The preparation of the noisy data has been described in the preceding section. HMM with 5 states and 64 mixture component performs best in the previous experiments, due to this only this best HMM model has been used to test the robustness of the new feature set. Phoneme classification accuracy was evaluated for SNRs in

Table 6. Classification results of different vowel class with desk mounted microphone (speaker independent)



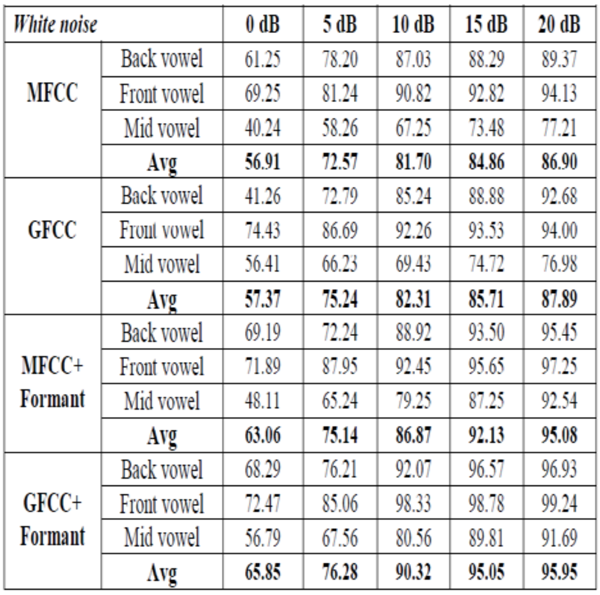
N(Observation State)		4	4	4	4	4	5	5	5	5	5
M(mixture)		**8**	**16**	**32**	**64**	**128**	**8**	**16**	**32**	**64**	**128**
MFCC	Back vowel	68.54	75.24	76.25	78.85	72.78	62.45	70.5	77.12	77.95	67.95
	Front vowel	66.48	69.15	70.3	72.95	69.88	72.15	71.78	75.42	75.75	71.47
	Mid vowel	39.78	41.06	43.09	45.95	46	41.6	42.85	46.87	48.8	46.5
	Avg	**58.27**	**61.82**	**63.21**	**65.92**	**62.89**	**58.73**	**61.71**	**66.47**	**67.50**	**61.97**
GFCC	Back vowel	68.13	74.36	74.92	78.27	74.32	58.42	69.95	79.12	80.95	73.28
	Front vowel	67.25	70.06	73.10	75.43	75.06	81.37	71.88	70.45	72.58	68.06
	Mid vowel	40.56	42.78	45.25	48.00	44.22	41.45	42.15	46.94	49.65	45.17
	Avg	**58.65**	**62.40**	**64.42**	**67.23**	**64.53**	**60.41**	**61.33**	**65.50**	**67.73**	**62.17**
MFCC+ Formant	Back vowel	69.58	77.00	79.97	82.68	75.00	71.25	72.35	77.50	82.15	77.22
	Front vowel	66.54	70.36	78.45	80.05	80.76	66.88	77.65	80.65	80.95	78.75
	Mid vowel	45.25	46.84	49.40	57.60	50.90	41.48	43.54	51.35	55.75	53.06
	Avg	**60.46**	**64.73**	**69.27**	**73.44**	**68.89**	**59.86**	**64.51**	**69.83**	**72.95**	**69.68**
GFCC+ Formant	Back vowel	68.69	72.15	78.05	80.50	75.25	69.45	70.48	78.42	81.45	79.15
	Front vowel	76.06	79.00	81.28	83.48	78.72	75.85	79.90	81.65	85.30	82.25
	Mid vowel	43.98	48.90	53.89	60.75	56.35	45.28	48.70	56.35	64.00	58.65
	Avg	**62.91**	**66.68**	**71.07**	**74.91**	**70.11**	**63.53**	**66.36**	**72.14**	**76.92**	**73.35**

Table 7. Classification efficiency of Hindi vowel with babble noise (speaker independent)



Babble Noise		0 dB	5 dB	10 dB	15 dB	20 dB
MFCC	Back vowel	65.25	83.45	87.19	89.73	90.05
	Front vowel	74.80	87.15	93.37	97.13	97.83
	Mid vowel	42.68	51.75	54.50	55.20	56.68
	Avg	**60.91**	**74.12**	**78.35**	**80.69**	**81.52**
GFCC	Back vowel	64.86	79.64	82.78	83.78	84.65
	Front vowel	72.50	88.80	91.23	91.38	91.53
	Mid vowel	51.32	73.96	75.66	77.17	78.74
	Avg	**62.89**	**80.80**	**83.22**	**84.11**	**84.97**
MFCC+ Formant	Back vowel	66.08	90.63	89.85	90.75	91.50
	Front vowel	76.55	83.49	92.48	92.94	93.85
	Mid vowel	57.54	80.75	84.75	86.45	89.83
	Avg	**66.72**	**84.96**	**89.03**	**90.05**	**91.73**
GFCC+ Formant	Back vowel	67.85	91.41	92.02	94.47	93.83
	Front vowel	78.95	86.67	93.48	96.09	96.54
	Mid vowel	62.45	80.96	90.28	94.35	95.30
	Avg	**69.75**	**86.35**	**91.93**	**94.97**	**95.22**

Table 8. Classification efficiency of Hindi vowel with factory noise (speaker independent)

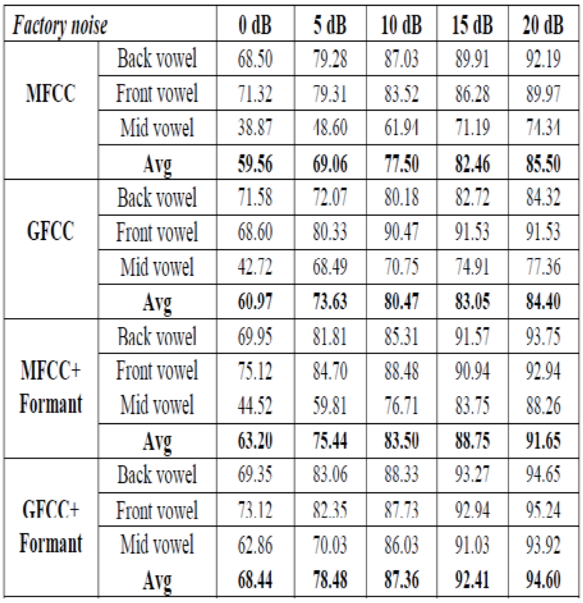


White noise		0 dB	5 dB	10 dB	15 dB	20 dB
MFCC	Back vowel	61.25	78.20	87.03	88.29	89.37
	Front vowel	69.25	81.24	90.82	92.82	94.13
	Mid vowel	40.24	58.26	67.25	73.48	77.21
	Avg	**56.91**	**72.57**	**81.70**	**84.86**	**86.90**
GFCC	Back vowel	41.26	72.79	85.24	88.88	92.68
	Front vowel	74.43	86.69	92.26	93.53	94.00
	Mid vowel	56.41	66.23	69.43	74.72	76.98
	Avg	**57.37**	**75.24**	**82.31**	**85.71**	**87.89**
MFCC+ Formant	Back vowel	69.19	72.24	88.92	93.50	95.45
	Front vowel	71.89	87.95	92.45	95.65	97.25
	Mid vowel	48.11	65.24	79.25	87.25	92.54
	Avg	**63.06**	**75.14**	**86.87**	**92.13**	**95.08**
GFCC+ Formant	Back vowel	68.29	76.21	92.07	96.57	96.93
	Front vowel	72.47	85.06	98.33	98.78	99.24
	Mid vowel	56.79	67.56	80.56	89.81	91.69
	Avg	**65.85**	**76.28**	**90.32**	**95.05**	**95.95**

the range of 0 dB to 20 dB. The classification efficiency of Hindi vowel recognizer under different level and type of noise is shown in Tables 7 to 9 respectively.

GFCC outperforms MFCC in most of the cases, because it takes the advantage of Gammatone filterbank which was designed according to the human cochlea. Furthermore, Formant frequency base features have shown better classification efficiency because they are considered to be less susceptible to channel distortion and noise. The accuracy achieved using the proposed features are found to be superior for all the phoneme classes, especially in mid vowel class. However, as noise level is getting high, the performance of the proposed features getting better. It can be seen from results at 0 dB and 5 dB SNR; the performance of proposed features is significantly better compared to

Table 9. Classification efficiency of Hindi vowel with white noise (speaker independent)

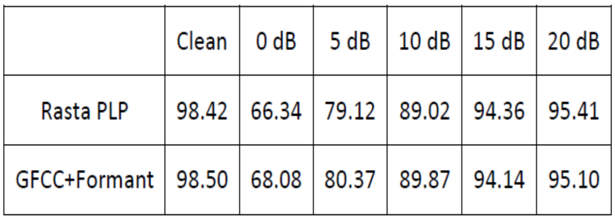


Factory noise		0 dB	5 dB	10 dB	15 dB	20 dB
MFCC	Back vowel	68.50	79.28	87.03	89.91	92.19
	Front vowel	71.32	79.31	83.52	86.28	89.97
	Mid vowel	38.87	48.60	61.94	71.19	74.34
	Avg	**59.56**	**69.06**	**77.50**	**82.46**	**85.50**
GFCC	Back vowel	71.58	72.07	80.18	82.72	84.32
	Front vowel	68.60	80.33	90.47	91.53	91.53
	Mid vowel	42.72	68.49	70.75	74.91	77.36
	Avg	**60.97**	**73.63**	**80.47**	**83.05**	**84.40**
MFCC+ Formant	Back vowel	69.95	81.81	85.31	91.57	93.75
	Front vowel	75.12	84.70	88.48	90.94	92.94
	Mid vowel	44.52	59.81	76.71	83.75	88.26
	Avg	**63.20**	**75.44**	**83.50**	**88.75**	**91.65**
GFCC+ Formant	Back vowel	69.35	83.06	88.33	93.27	94.65
	Front vowel	73.12	82.35	87.73	92.94	95.24
	Mid vowel	62.86	70.03	86.03	91.03	93.92
	Avg	**68.44**	**78.48**	**87.36**	**92.41**	**94.60**

conventional cepestral based feature. This shows higher robustness of the proposed features towards in the presence of high value of noise. It is also notable that, in most of the cases at low SNR level, cepstral based feature shows poor classification efficiency to classify mid vowel and back vowel class. Because the back vowel is formed by placing the tongue as far back as possible in the mouth without creating any constriction that would be classified as a consonant. Mid-vowel is articulated in the near of central cavity, and the jaw is approximately in the midway of its vertical motion that would be classified as a stop. The proposed feature have shown great improvement to classify the same class by taking advantage of formant frequencies. GFCC based formant feature has shown their supremacy over other features at low SNR level by taking the advantage of both auditory filterbank and robustness of formant frequencies.

Additionally to compare the performance of proposed feature with another robust feature Rasta-PLP has been taken. Table 10 shows the comparative avg. phoneme classification performance. Both shows nearly equal classification rate with clean data. Proposed feature have outperformed Rasta-PLP especially in the presence of high noise.

Table 10. Comparative classification performance (speaker independent)



	Clean	0 dB	5 dB	10 dB	15 dB	20 dB
Rasta PLP	98.42	66.34	79.12	89.02	94.36	95.41
GFCC+Formant	98.50	68.08	80.37	89.87	94.14	95.10

6 Conclusion

In this paper, we have presented a simple method for recognizing the vowels of the Hindi language in continuous speech. The proposed method is based on recognition of frequencies of first three formants that are present in vowels along with their cepstral feature. Performance of the proposed feature has been tested in clean condition, sensor mismatch condition and as well as three different noisy condition. Different type of HMM with the different number of states and mixture components are used to select the optimum one suited for our system. Most of the time, auditory based feature GFCC outperforms MFCC, especially across noisy condition because auditory perception based filterbank can focus and follow the target speech from the complex auditory spectrum. When compared to cepstral based feature, the formant based features along with cepstral features have performed better under all noise conditions and as well as sensor mismatch condition at the cost of slightly higher computational resources. Formant based GFCC has shown best classification efficiency across all testing conditions. We have demonstrated the effectiveness of proposed feature on 90 speaker continuous Hindi speech database but it is required to study the performance of proposed feature for continuous phoneme recognition. The performance of proposed feature should be studied for consonants to develop a robust Hindi ASR.

Acknowledgements: We are thankful to the respected reviewers and honourable editor for providing important suggestions and constructive comments which have helped us improve the quality of the paper.

References:

[1] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust. Speech Signal Process*, vol. ASSP-28, pp. 357–366, 1980.

[2] E. Wong and S. Sridharan, “Comparison of Linear Prediction Cepstrum Coefficients and Mel-Frequency Cepstrum Coefficients for Language Identification,” in *In Proceedings of IEEE International Symposium on*

Intelligent Multimedia Video and Speech Processing., 2001, pp. 95–98.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.

[5] A. Biswas, P. K. Sahu, and M. Chandra, "Admissible Wavelet Packet Features based on Human Inner Ear Frequency Response for Hindi Consonant Recognition," *Comput. Electr. Eng.*, vol. Accepted, 2014.

[6] D. S. Mary, D. Manimegalai, and B. G. Ram, "Wavelets and Ridgelets for Biomedical Image Denoising," *WSEAS Trans. Syst.*, vol. 12, no. 3, pp. 165–178, 2013.

[7] P. Talukder, M. Sarma, and K. K. Sarma, "Recognition of Assamese SpokenWords using a Hybrid Neural Framework and Clustering Aided Apriori Knowledge," *WSEAS Trans. Syst.*, vol. 12, no. 7, pp. 360–370, 2013.

[8] A. N. Mishra, M. C. Sharan, A. Biswas, and S. N. Sharan, "Hindi phoneme-viseme recognition from continuous speech," *Int. J. Signal Imaging Syst. Eng.*, vol. 6, no. 3, pp. 164–171, 2013.

[9] K. Samudravijaya, "Durational characteristics of Hindi stop consonants," in *Proc. of Eurospeech, Geneva*, 2003, pp. 81–84.

[10] M. Kumar, N. Rajput, and A. Verma, "A large-vocabulary continuous speech recognition system for Hindi," *IBM J. Res. Dev.*, vol. 48, no. 5.6, pp. 703–715, 2004.

[11] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *the proceedings of IEEE Acoustics, Speech and Signal Processing*, 2009, pp. 4625–4628.

[12] Z. Tuske, P. Golik, R. Schluter, and F. R. Drepper, "Non-stationary feature extraction for automatic speech recognition," in *IEEE Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5204–5207.

[13] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Comput. Speech Lang. Elsevier*, vol. 24, no. 1, pp. 77–93, 2010.

[14] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, *An efficient auditory filterbank based on the Gammatone function*. Appl. Psychol. Unit, Cambridge University, 1988.

[15] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. Vol. 109, pp. 1101–1109, 2001.

[16] A. S. Bregman, *Auditory Scene Analysis*. Cambridge . MIT Press, 1990.

[17] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications,*. Wiley interscience, 2006.

[18] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA.: PTR Prentice-Hall, 1993.

[19] B. Chen and P. C. Loizou, "Formant frequency estimation in noise," in *In IEEE Proceedings of Acoustics, Speech, and Signal Processing*, pp. 581–584.

[20] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 129–134, 1993.

[21] F. Neri, "Quantitative Estimation of Market Sentiment: a discussion of two alternatives," *WSEAS Trans. Syst.*, vol. 11, no. 12, pp. 691–702, 2012.

[22] Z. Bojkovic and F. Neri, "An introduction to the special issue on advances on interactive multimedia systems," *WSEAS Trans. Syst.*, vol. 12, no. 7, pp. 337–338, 2013.

[23] P. Hájek and F. Neri, "An introduction to the special issue on computational techniques for trading systems, time series forecasting, stock market modeling, financial assets modeling," *WSEAS Trans. Syst.*, vol. 10, no. 4, pp. 291–292, 2013.

[24] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1312–1322, 2006.

[25] L. Welling and H. Ney, “Formant estimation for speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, 1998.

[26] H. R. Iqbal., M. M. Awais, S. Masud, and S. Shamail, “On vowels segmentation and identification using formant transitions in continuous recitation of Quranic Arabic,” in *In New challenges in applied intelligence technologies(Springer)*, 2008, pp. 155–162.

[27] Y. A. Alotaibi and A. Hussain, “Comparative analysis of Arabic vowels using formants and an automatic speech recognition system,” *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 3, no. 2, pp. 11–21, 2010.

[28] G. N. Kodandaramaiah, M. N. Giriprasad, and M. M. Rao, “Independent speaker recognition for native English vowels,” *Int. J. Electron. Eng. Res.*, vol. 2, no. 3, pp. 377–381, 2010.

[29] D. A. Kocharov, “Automatic vowel recognition in fluent speech,” in *In Proceedings of the 9th Conference of Speech and Computer*.

[30] Q. Yan, S. Vaseghi, E. Zavarehei, B. Milner, J. Darch, P. White, and I. Andrianakis, “Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing,” *Comput. Speech Lang.*, vol. 21, no. 3, pp. 543–561, 2007.

[31] K. Samudravijaya, P. V. S. Rao, and S. S. Agrawal, “Hindi speech database. In Proc. China; October 2002:p. 456–9.,” in *Int. Conf. on Spoken Language processing (ICSLP00) Beijing*, 2002, pp. 456–459.

[32] K. F. Lee and H. W. Hon, “Speaker-independent phone recognition using hidden Markov models,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 14, pp. 1641–1648, 1989.

[33] A. N. Mishra, M. Chandra, A. Biswas, and S. N. Sharan, “Robust features for connected Hindi digits recognition,” *Int. J. Signal Process. Image Process. Pattern Recognition,*, vol. 4, no. 2, pp. 79–90, 2011.