Wireless Personal Communications (2021) 120:3225–3242
https://doi.org/10.1007/s11277-021-08610-0

Speech Recognition Using Enhanced Features with Deep Belief Network for Real Time Application

Gurpreet Kaur[1] · **Mohit Srivastava**[2] · **Amod Kumar**[3]

Accepted: 7 June 2021 / Published online: 16 June 2021

Abstract
Speech is a natural way used by humans to communicate information. Speech signal conveys the linguistic information (the message) and lot of information about the speaker himself: Gender, age, regional origin, health and emotional state. Speech recognition is the technology of letting a machine understand human speech. Speaker recognition is the technology by which a machine distinguishes different speakers from each other. In real life, speaker and speech recognition have been used very frequently which vary from healthcare, military to applications pertaining to daily use. These may include but are not limited to commanding electronic devices through speech. All existing speech recognition system perform efficiently in control environment. But for real time applications, the performance gets affected because of the variability in speaking style and background noise. In order to deal with all these effects, enhanced Mel Frequency Cepstral Coefficients (MFCC) are calculated. Deep belief network (DBN) of stacked restricted Boltzmann machine (RBM) is used for training and testing. The proposed system is implemented using standard TIDIGITS dataset giving 97.29% accuracy.

Keywords Speaker recognition · Speech recognition · Mel frequency cepstral coefficients (MFCC) · Principal component analysis (PCA) · Deep belief network (DBN)

✉ Gurpreet Kaur
regs4gurpreet@yahoo.co.in

✉ Mohit Srivastava
mohitsrivastava.78@gmail.com

✉ Amod Kumar
csioamod@yahoo.com

[1] University Institute of Engineering & Technology, Panjab University, Chandigarh 160025, India

[2] Chandigarh Engineering College, Landran, Mohali 140307, India

[3] National Institute of Technical Teachers Training and Research, Chandigarh 160019, India

1 Introduction

Speech signal conveys many levels of information. At primary level, speech signal gives the message; at secondary level, speech signal conveys the information about speaker. Speech production is very complex phenomenon. It comprises of many levels of processing. First of all, message planning is done in our mind and then language coding is done. Based upon the coding, neuromuscular command is generated. After this, sound is produced through vocal cords. Every human speech is different from each other because of different parameters like linguistics (lexical, syntactic, semantic, pragmatics), para linguistics (intentional, attitudinal, stylistic) and non-linguistics (physical, emotional). Therefore, speech signal contains different segmental and supra segmental features which can be extracted for the speaker as well as speech recognition [1]. Speech signals are considered as highly non-stationary signals as they are not only difficult to observe but also more prone to noise. A number of feature extraction techniques have been developed whose primary focus is to preserve the variations along with relevant speech signal information that is compact as well as reasonable. On the basis of biological classification, features can be derived into two parts such as production as well as perception based. On the basis of domain of processing, they can be divided into temporal, eigen, cepstral and frequency domains. The temporal features included amplitude, power and zero crossing rate which are used to take voicing decisions. Other features such as brightness, tonality, loudness as well as pitch are considered under perceptual features. In speech recognition systems, detecting the presence of speech in a noisy environment is considered as a crucial problem in case of real time applications. Speech recognition is a wide field in terms of different feature extraction techniques, recording environment, databases and classification methods. There are many feature extraction techniques like linear predictive coding coefficient (LPCC), perceptual linear prediction (PLP), relative spectra filtering (RASTA) and Mel frequency cepstral coefficients (MFCC). There are various types of databases available for speech and speaker recognition systems from isolated to continuous speech like TIDIGITS, RSR 2015, TIMIT etc. There are basically three approaches for matching/modeling; acoustic phonetic approach, pattern recognition approach (dynamic time warping, Hidden Markov model, vector quantization) and artificial Intelligence approach (neural network, deep neural network). In this paper, combined speaker and speech recognition system is proposed using enhanced MFCC features with DBN for real time applications.

2 Related Work

Research in speaker and speech recognition field has made tremendous progress in the last 60 years. Speech recognition systems can be divided into four generations. In 1st generation (1950s–1960s), the work was based upon the acoustic phonetics approaches. Template matching techniques like LPC, DTW etc. was used in 2nd generation (1960s–1970s). In 3rd generation (1970s–2000s), statistical modeling techniques like HMM were mostly used by the researchers. However, in the current age also known as 4th generation (2000s onwards), the focus is on deep learning [2–5]. Nascimento, T.P., et al., 2011 [6] have proposed a speech recognition system using ANN and HMM for English words. The recognition rate achieved for HMM is 96% and for ANN is 97%. In this paper, adaptive learning rate and large dataset have increased the recognition rate. Guojiang, F., 2011 [7] has

implemented the system using two types of classifiers; multiple layer perceptron (MLP) and radial basis function (RBF). LPC Coefficients are used as features. RBF classifier performance is superior than MLP for 16 LPCC speaker dependent coefficients. Seyedin, S., et al. 2013 [8] have proposed a new type of feature based upon MVDR spectrum of filtered autocorrelation sequence. They have used TIDIGITS database for the implementation and achieved 76.6% of accuracy. Initially there were many problems faced by researchers in training hidden layers due to propagating training errors to hidden layers and getting stuck in local minima etc. [9, 10]. But in 2006, G. Hinton et al. [11] introduced Deep Belief Network (DBN), with layer wise training. Many researchers have used DBN with supervised or unsupervised training algorithms [12–14]. Literature reported that systems based on DBN are better than HMM, GMM, GMM-HMM techniques [15–22]. Jaitly, N. et al. 2011[23] used DBN using Restricted Boltzmann Machine (RBM) and contrastive divergence (CD) training algorithm on speech signals. They have tested on TIMIT corpus and achieved better performance in terms of position independent word error rate (PER) [24, 25]. Then this work was carried forward by Mohamed, A. et al. [26] and they have used MFCC features with DBN to improve performance of 20.7% PER as compared to using raw speech features. Dhanashri, D., and Dhonde, S.B., 2016 [27] have used HMM for acoustic modeling and DBN is used as classifier. They have used TIDIGITS as database and achieved 96.58% accuracy.

The speech recognition systems have been deployed in various domains such as for domestic use, educational purpose, for the purpose of entertainment, in medical science for healthcare and medical transcriptions etc. It is observed that less work has been done for rehabilitation application of speech recognition system [28]. Thus, this study aims to develop an application-oriented research work for handicapped persons for combined speaker and speech recognition for any activities like voice operated wheelchair.

3 Proposed Work

All existing speech recognition system perform efficiently for stored database. But for real time applications, the performance gets affected because of the variability in speaking style and background noise. In order to deal with all these effects, enhanced MFCC features are calculated. In this, two types of variations are calculated that can be there in generalized features of MFCC, named as tolerance 1 and tolerance 2. The features which are used in the proposed model are fusion of two-level mathematical analysis of the feature extraction method. Features are calculated in three phases: calculation of tolerance1, calculation of tolerance 2 and PCA fusion.

3.1 Calculation of Tolerance1

Despite the recording of samples in same environment and with same speakers, variations in samples were observed, due to intra speaker variability. TIDIGITS dataset is used which is available for eleven isolated words (zero, one, two, three, four, five…ten) for 326 speakers. First of all, speech signal is converted into frequency domain to know about the frequencies of the speech signal using FFT. The output of FFT contains a lot of data that is not required because at higher frequencies there is not any difference between the frequencies. This is based upon the phenomenon of human hearing. The scale is linear until 1000 and logarithmic after it. So, to calculate the energy level at each frequency, MEL scale

analysis is done using MEL filters. Then energy is calculated. After that, logarithmic of filter bank energies is taken. This operation is done to match the features closer to human hearing. At last, DCT of the log filter bank energies is taken to decorrelate the overlapped energies. First 13 coefficients are selected known as MFCC features. This is because higher features degrade the recognition accuracy of the system. Those features don't carry speaker and speech related information.

Due to variations in all words, the feature values are also different for every word, in spite of being spoken by same speaker. The difference between all samples of same word is calculated and is represented as shown in Eq. (1).

$$D_{ij}^{K} = \sqrt{\left(X_i - X_j\right)^2}, 1 \le i,j \le n, \tag{1}$$

where n is the no. of samples of each word, X_i and X_j are voice samples, K = Number of isolated words.

Hence distance matrix for all words is calculated as represented in Eq. (2).

$$Dist^k = \begin{matrix} D_{11} \\ \cdot \\ \cdot \\ D_{ij} \end{matrix} \tag{2}$$

where i and j varies with no. of samples.

Then maximum and minimum values are calculated as $\max\left(D_{ij}\right) and\ min\left(D_{ij}\right)$ After this, variations are calculated by subtracting minimum value from its maximum value for each word as represented in Eq. (3).

$$Var^k = \left[max\left(D_{ij}\right) - min\left(D_{ij}\right)\right] \tag{3}$$

After this, mean of all samples of every word is calculated as given in Eq. (4).

$$M^k = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{4}$$

where X_i is the voice sample of each word. This value is subtracted from the mean value of each sample and this is called tolerance 1 as shown in Eq. (5).

$$Tol\ 1^k = M^k - Var^k \tag{5}$$

3.2 Calculation of Tolerance 2

In second step, mean variations are calculated. This is called tolerance 2. For calculating the tolerance 2, instead of taking differences between the individual samples, here difference between mean value and samples of same word is calculated as represented in Eq. (6).

$$MD_{ij}^{K} = \sqrt{\left(M_{ij} - X_{ij}\right)} \tag{6}$$

where M_{ij} = mean of word, X_{ij} = voice sample.

Rest of the procedure is same as tolerance 1 is calculated as explained above from Eqs. (2) to (5).

3.3 PCA Fusion

Now there are features calculated from tolerance 1 and tolerance 2 method and hence total 26 features are there for each word. So, to decide which features are selected out of tolerance 1 and tolerance 2 features, principal component analysis (PCA) is used [29]. This algorithm is based upon how much amount of the tolerance 1 features and tolerance 2 features will be taken to get final tolerance features. It is a mathematical based procedure that involves the transformation of features into principal components that computes a reduced and important feature set. Figure 1is showing the flow chart for the PCA fusion process.

3.4 Deep Neural Network

In deep neural network there are many hidden layers in it. The base is taken from visual cortex. The brain processes the information through several sections of brain. The neurons in each section behave differently. So neural network can be modeled as multilayer network consisting of lower level to higher level of features [30–32]. Training is the major issue in deep neural networks because optimization is difficult. There may be under-fitting and over-fitting in the system. Under-fitting is due to vanishing gradient problem and over-fitting is because of high variance and low bias situation. There is one solution for this, that is unsupervised pre training approach. Unsupervised pre training is done one layer at a time. Features are fed to first hidden layer, then second layer takes the combinations of features from first layer. This process goes on till the last layer. After that, supervised training is done for entire network.

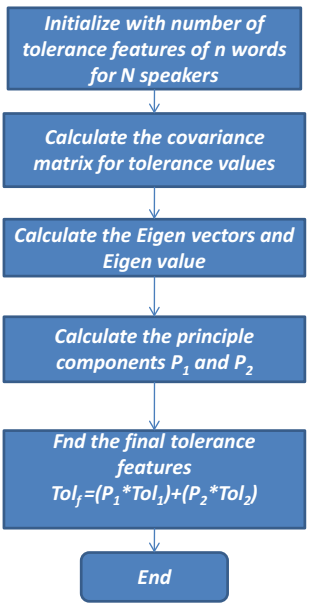

Fig. 1 PCA fusion

4 Results and Discussion

In this research work, standard (TIDIGITS) database is used. TIDIGITS is available for eleven isolated words (zero, one, two, three, four, five…ten) for 326 speakers. Total 2260 isolated words are taken, spoken by 57 women and 56 men speakers. First of all, MFCC features are calculated. Following Table 1 is showing the MFCC feature coefficients (Cf.1 to Cf.13) of speaker 1 for first five words i.e., 'ONE', 'TWO', 'THREE', 'FOUR', 'FIVE'. The results are shown for five words of one speaker.

When same speaker uttered same words then also there is variability in speech signal. The following Table 2 shows the variations in MFCC features when same set of words are spoken by same speaker 1.

To deal with these variations enhanced MFCC features are calculated. Firstly, tolerance 1 is calculated which is the variations of the single word with other words spoken by the same speaker. Therefore, difference between all samples of same word is calculated. Table 3 shows the difference of MFCC features for word set 1 and word set 2.

Above results shows the distance matrix (D_i^K) represented as in Eq. 1. There are five words, so five distances are found out. Then maximum and minimum values are calculated from each distance matrix represented as max (D_i) and min (D_i). In Table 3 maximum values are represented in italic and minimum values are represented in bold.

Then variation is calculated by subtracting minimum value from its maximum value and represented as Var^k as per Eq. 3. The values for variations for all words is shown in Table 4.

After this, mean (M^k) of all sample words is calculated as shown in Table 5.

Then tolerance 1 is calculated by subtracting variations from its mean value as shown in Table 6.

Hence based upon the number of speakers, tolerance values can be calculated. In second scenario for calculating the tolerance 2, mean variations are calculated instead of words variations. In this method, firstly the difference between mean value and sample value is calculated as per Eq. 6. These difference values depend upon the number of samples for each word. Following Table 7 shows the tolerance 2 features.

Now there are features for tolerance 1 and tolerance 2 and total 26 values are there for each word. So PCA algorithm is used to fuse these values to get 13 important features. In PCA algorithm firstly, covariance matrix is found out as shown below for the word ONE.

$$C = \begin{bmatrix} 272.2445 & 272.4905 \\ 272.4905 & 272.8740 \end{bmatrix}$$

Eigen values and Eigen vectors are calculated for the word ONE as shown below.

$$Eigen\ Value = \begin{bmatrix} 0.0686 & 0 \\ 0 & 545.0499 \end{bmatrix}$$

$$Eigen\ Vector = \begin{bmatrix} -0.7075 & 0.7067 \\ 0.7067 & 0.7075 \end{bmatrix}$$

Then principal components $P_1 = 0.4993$ and $P_2 = 0.5003$ are calculated. Similarly, all these steps are repeated to get the principal components for all words. Final fused tolerance features are shown in Table 8 for all words.

Table 1 MFCC features of speaker 1 for first set of words

WORD SET 1	MFCC Features of speaker 1												
	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
ONE	57.98	3.00	3.63	5.92	−5.86	−1.65	−9.45	6.39	2.77	−0.23	0.04	0.28	0.59
TWO	56.72	−1.94	3.10	5.95	−0.42	2.04	2.73	3.58	3.11	2.16	−0.30	4.06	−0.46
THREE	59.43	−4.41	3.15	9.65	0.48	6.72	−3.55	6.46	−2.50	4.05	0.32	1.95	−0.13
FOUR	61.94	8.32	2.23	−1.61	−4.20	0.10	−6.63	2.73	−1.25	0.52	−0.41	1.68	−1.65
FIVE	54.27	−9.07	2.10	2.91	3.42	−3.95	−0.94	0.34	−0.70	−4.30	−0.77	0.54	−2.13

Table 2 MFCC features of speaker 1 for second set of words

WORD SET 2	MFCC Features of speaker 1												
	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
ONE	57.80	3.48	3.07	6.29	−6.66	1.43	−10.64	8.17	2.21	2.52	0.12	2.94	0.16
TWO	60.22	−0.75	2.35	6.33	1.25	1.62	3.03	2.69	1.63	2.43	−3.14	4.75	−1.08
THREE	61.16	−4.03	3.15	9.53	2.03	7.73	−3.78	7.22	−2.60	3.49	0.76	1.47	−0.35
FOUR	56.43	6.42	1.91	0.11	−1.23	−1.17	−5.57	−0.14	0.95	−0.74	0.71	0.12	−1.36
FIVE	53.28	−9.38	3.17	2.87	3.51	−5.35	−0.65	1.04	−0.89	−3.55	−1.96	0.92	−1.69

Table 3 Difference of MFCC features of speaker 1

Difference of MFCC Features for speaker1													
Difference (D^K)	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
D_1^{ONE}	0.18	0.47	0.56	0.36	0.80	*3.09*	1.18	1.77	0.56	2.75	**0.07**	2.66	0.43
D_2^{TWO}	*3.49*	1.18	0.74	0.37	1.67	0.41	0.30	0.88	1.48	**0.26**	2.84	0.69	0.61
D_3^{THREE}	*1.73*	0.38	**0.00**	0.12	1.55	1.00	0.23	0.75	0.09	0.56	0.44	0.47	0.21
D_4^{FOUR}	*5.51*	1.89	0.31	1.73	2.96	1.28	1.06	2.87	2.21	1.27	1.12	1.81	**0.28**
D_5^{FIVE}	0.98	0.31	1.07	**0.04**	0.09	*1.39*	0.28	0.70	0.18	0.74	1.19	0.38	0.43

Bold values are the minimum values, that are calculated from distance matrix represented by Eq. (2)

Table 4 Variations in words

Words	$Var^k = \max(d_i) - \min(d_i)$
Var^{ONE}	3.01
Var^{TWO}	3.22
Var^{THREE}	1.73
Var^{FOUR}	5.22
Var^{FIVE}	1.35

Now there are 13 final enhanced features for every word. The mean of MFCC features and enhanced MFCC features is taken and after normalization these are fed to deep belief network for training. Two hidden layers with 200 and 300 neurons are used in DBN with Contrastive Divergence learning rule for the 20 epochs. Final DBN input is shown in Table 9 for speaker 1.

Similarly, enhanced features are calculated for all speakers and data is fed to DBN for training.

A comparison is done between MFCC features and enhanced MFCC features. The results showed that when system is trained with enhanced MFCC features, the accuracy is much better at different SNRs as compared to the baseline MFCC features. The reason is that enhanced MFCC features are calculated by taking care of the variation in speech signal. It may be due to the intra speaker variability or due to the environmental effects. For clean speech signal the accuracy is about 94% for MFCC features and 97% is for enhanced MFCC features. At 15 dB, when MFCC features are used the accuracy is 56% and when enhanced MFCC features are used, the accuracy is 96%. It showed that accuracy decreases about half when there is variability in speech signals using MFCC features, but for enhanced MFCC trained DBN, the accuracy is almost same as for clean signals. This showed that system trained with enhanced MFCC features achieved good accuracy even in noisy conditions as shown in Table 10.

A comparison is done between proposed system and baseline system on standard dataset (TIDIGITS). Seyedin Sanaz et.al [8] have used features which are computed from minimum variance distortion less response (MVDR) spectrum by modifying the PLP technique known as PMSR features. In this weighting of sub band is modified to get MVDR spectrum and then LP coefficients are transformed to get robust PMSR

Table 5 Mean of MFCC features for speaker 1

Mean of MFCC Features for speaker1													
Mean (M^k)	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
M_1^{ONE}	57.89	3.24	3.35	6.10	−6.26	−0.10	−10.04	7.28	2.49	1.14	0.08	1.61	0.37
M_2^{TWO}	58.47	−1.35	2.72	6.14	0.41	1.83	2.88	3.14	2.36	2.29	−1.72	4.41	−0.77
M_3^{THREE}	60.30	−4.22	3.15	9.59	1.25	7.23	−3.66	6.84	−2.55	3.77	0.54	1.71	−0.24
M_4^{FOUR}	59.18	7.37	2.07	−0.74	−2.71	−0.53	−6.10	1.29	−0.15	−0.10	0.14	0.77	−1.50
M_5^{FIVE}	53.77	−9.22	2.64	2.89	3.47	−4.65	−0.80	0.69	−0.79	−3.92	−1.37	0.73	−1.91

Table 6 Tolerance 1 for speaker 1

WORDS	Tolerance 1 ($Tol^k = M^k\text{-}Var^k$) for speaker1												
Tolerance (T^k)	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
T_1^{ONE}	54.84	0.22	0.33	3.08	−9.28	−3.12	−13.06	4.26	−0.52	−1.87	−2.93	−1.40	−2.64
T_1^{TWO}	55.24	−4.58	−0.50	2.91	−2.81	−1.39	−0.34	−0.08	−0.85	−0.93	−4.94	1.18	−4.00
T_1^{THREE}	58.57	−5.95	1.42	7.86	−0.46	5.50	−5.39	5.11	−4.28	2.05	−1.18	−0.01	−1.97
T_1^{FOUR}	53.96	2.14	−3.15	−5.97	−7.94	−5.76	−11.33	−3.93	−5.37	−5.33	−5.02	−4.44	−6.73
T_1^{FIVE}	52.42	−10.5	1.29	1.54	2.12	−6.0	−2.15	−0.65	−2.15	−5.27	−2.72	−0.61	−3.26

Table 7 Tolerance 2 for speaker 1

Tolerance 2 (T^k)	Tolerance 2 ($Tol^k = M^k\text{-}Var^k$) for speaker1												
	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
T_2^{ONE}	56.42	1.61	1.98	4.50	−7.57	−2.39	−11.26	5.33	1.12	−1.05	−1.44	−0.55	−1.02
T_2^{TWO}	55.98	−3.26	1.30	4.43	−1.62	0.32	1.19	1.74	1.13	0.61	−2.62	2.62	−2.23
T_2^{THREE}	59.00	−5.18	2.29	8.76	0.00	6.11	−4.47	5.79	−3.39	3.05	−0.42	0.97	−1.05
T_2^{FOUR}	57.95	5.23	−0.46	−3.79	−6.06	−2.82	−8.98	−0.60	−3.31	−2.40	−2.75	−1.38	−4.19
T_2^{FIVE}	53.35	−9.82	1.69	2.22	2.77	−4.98	−1.15	−0.15	−1.42	−4.79	−1.74	−0.03	−2.69

Table 8 Fused tolerance features for speaker 1

Final Tolerance (T_f^k)	$T_f = (P_1 * T_1) + (P_2 * Tol_2)$ for speaker1												
	Cf.1	Cf.2	Cf.3	Cf.4	Cf.5	Cf.6	Cf.7	Cf.8	Cf.9	Cf.10	Cf.11	Cf.12	Cf.13
T_{ONE}	55.64	0.92	1.16	3.79	−8.43	−2.76	−12.16	4.80	0.29	−1.46	−2.18	−0.98	−1.83
T_{TWO}	55.61	−3.92	0.39	3.66	−2.22	−0.54	0.41	0.82	0.13	−0.16	−3.79	1.89	−3.12
T_{THREE}	58.78	−5.57	1.85	8.31	−0.23	5.81	−4.93	5.45	−3.83	2.55	−0.80	0.47	−1.51
T_{FOUR}	55.98	3.71	−1.79	−4.87	−6.99	−4.27	−10.14	−2.24	−4.33	−3.85	−3.90	−2.89	−5.45
T_{FIVE}	52.89	−10.2	1.49	1.88	2.44	−5.49	−1.85	−0.40	−1.78	−5.03	−2.23	−0.32	−2.98

Table 9 Input to DBN for speaker 1

0.9660	0.0351	0.0381	0.0839	−0.1255	−0.0249	−0.1894	0.1025	0.0234	−0.0032	−0.0183	0.0050	−0.0127
0.9661	−0.0452	0.0260	0.0826	−0.0159	0.0104	0.0275	0.0331	0.0208	0.0176	−0.0472	0.0529	−0.0335
0.9675	−0.0797	0.0406	0.1454	0.0082	0.1059	−0.0701	0.0998	−0.0521	0.0513	−0.0023	0.0177	−0.0145
0.9664	0.0923	0.0015	−0.0481	−0.0826	−0.0412	−0.1374	−0.0088	−0.0386	−0.0341	−0.0325	−0.0186	−0.0593
0.9686	−0.0516	−0.1127	−0.0157	−0.1343	−0.0109	−0.1347	−0.0332	0.0268	−0.0730	0.0208	−0.0261	−0.0175
0.9639	−0.01757	0.0373	0.0431	0.0534	−0.0917	−0.0241	0.0025	−0.0235	−0.0811	−0.0327	0.0036	−0.0443

Table 10 Comparison of % accuracy between MFCC and enhanced MFCC features on TIDIGITS at different SNR

SNR (dB)	% Accuracy on TIDIGITS Database using MFCC features	% Accuracy on TIDIGITS Database using enhanced MFCC features
Clean	94.59	97.29
20	62.16	97.29
15	56.75	96.09
10	54.05	91.89
5	54.05	83.78
0	51.35	75.67
− 5	35.13	59.45

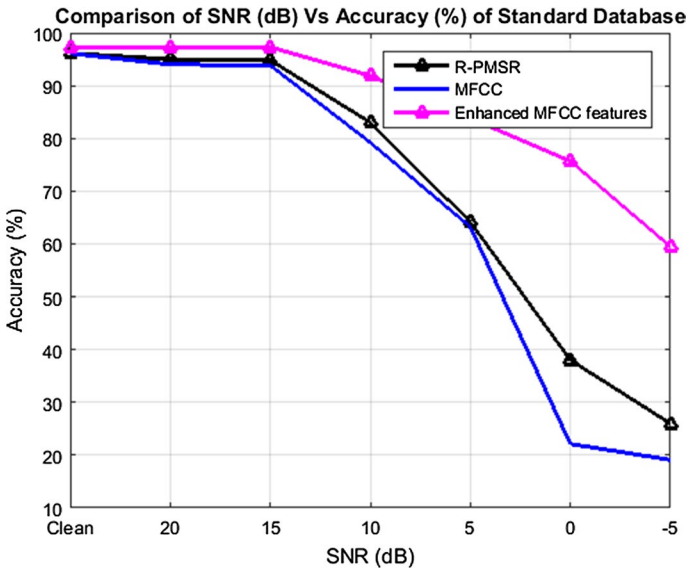

Fig. 2 Comparison of % accuracy with SNR

(R-PMSR) features. A comparison is done between the R-PMSR, MFCC and proposed system using enhanced MFCC features at different SNRs as shown in Fig. 2.

The results showed that proposed system has achieved better accuracy at different SNRs when compared with baseline systems. As in baseline systems, the accuracy decreases sharply with high noise environment. But enhanced MFCC features worked well in noisy environment. Therefore, for real time applications where both intra speaker as well as environment effects are of greatest interest, enhanced MFCC features are the best.

5 Conclusion

In proposed work, enhanced MFCC features are calculated in terms of tolerance 1 and tolerance 2. The recognition accuracy improves around 3% when enhanced MFCC features are used as compared to baseline MFCC features. Experimentation is done on TIDIGITS. Deep belief network is used for the classification purpose. Comparison is done between proposed technique and existing techniques. Baseline system using MFCC features has given 94.59% accuracy, R-PMSR feature based system has given 95.12% accuracy and enhanced MFCC based system has given 97.29% accuracy.

Availability of data and material TIDIGITS dataset is an open source.

Code availability Readers may ask.

Declaration

Conflicts of interest The authors declare that they have no conflict of interest.

Ethics approval Punjab University, Chandigarh has given the permission to do research work**.**

Consent for publication The authors have given the consent for publication.

References

1. Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Hoboken: Prentice-hall publishers.
2. Siniscalchi, S. M., Svendsen, T., & Lee, C. H. (2014). An artificial neural network approach to automatic speech processing. *Neurocomputing, 140*, 326–338.
3. Dede, G., and Sazlı, M. H. (2015). "Speech recognition with artificial neural networks," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 20(3), 763–768.
4. Richardson, F., Member, S., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters, 22*(10), 1671–1675.
5. Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication, 120*, 11–19.
6. Nascimento, T. P., & Stefanoy, D. (2011). Speech Recognition using Artificial Neural Networks. *Simposio Brasileiro de Automaçao Inteligente, 10*, 1316–1321.
7. Guojiang, F. (2011) "A novel isolated speech recognition method based on neural network". 2nd International Conference on Networking and Information Technology, Singapore, 17, 64–69.
8. Seyedin, S., Mohammad, A., and Gazor, S. (2011) "New features using robust MVDR spectrum of filtered autocorrelation sequence for robust speech recognition". *The Scientific World Journal*, p.1–11.
9. Salakhutdinov, R., & Hinton, G. (2012). An efficient learning procedure for deep Boltzmann machines. *Neural Computation, 24*(8), 1967–2006.
10. Huang, X., & Deng, L. (2010). Handbook of natural language processing. *An Overview of Modern Speech Recognition., second edition* (pp. 339–367). CRC publishers.
11. Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). Fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.
12. Keyvanrad, M. A., & Homayounpour, M. M. (2014) "A brief survey on deep belief networks and introducing a new object-oriented MATLAB toolbox (DeeBNet)", p. 1–25.
13. Le Cun, Y., Yoshua, B., & Geoffrey, H. (2015). Deep learning. *Nature, 521*(7553), 436–444.
14. Cai, M., & Liu, J. (2016). Maxout neurons for deep convolutional and LSTM neural networks in speech recognition. *Speech Communication, 77*, 53–64.

15. Dighe, P., Asaei, A., & Bourlard, H. (2016). Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition. *Speech Communication, 76*, 230–244.
16. Sarikaya, R., Hinton, G., & Deoras, A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(4), 778–784. https://doi.org/10.1109/TASLP.2014.2303296.
17. Mirsamadi, S., and Hansen, J. (2015) "A study on deep neural network acoustic model adaptation for robust far-field speech recognition," *Proceedings of Interspeech*, pp. 2430–2434.
18. Cutajar, M., Micallef, J., Casha, O., Grech, I., & Gatt, E. (2013). Comparative study of automatic speech recognition techniques. *IET Signal Processing, 7*(1), 25–46.
19. Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep Recurrent Neural Network. *IEEE International Conference, 3*, 6645–6649.
20. Bourouba, E. H., Bedda, M., & Djemili, R. (2006). Isolated words recognition system based on hybrid approach DTW/GHMM. *Informatica, 30*(3), 373–384.
21. Deng, L., Hinton, G., and Kingsbury, B. (2013) "New types of deep neural network learning for speech recognition and related applications: an overview". IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8599–8603.
22. Chandra, B., & Sharma, R. K. (2016). Fast learning in deep neural networks. *Neurocomputing, 171*, 1205–1215.
23. Jaitly, N., and Hinton, G. E. (2011) "Learning a better representation of speech sound waves using Restricted Boltzmann Machines". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5887.
24. Tanaka, M., and Okutomi, M., (2014), "A Novel Inference of a Restricted Boltzmann Machine". Proceedings - International Conference on Pattern Recognition, pp. 1526–1531.
25. Farahat, M., & Halavati, R. (2016). Noise robust speech recognition using deep belief networks. *International Journal of Computational Intelligence and Applications, 15*(1), 1–17.
26. Mohamed, A., Dahl, G., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transaction on Audio, Speech and Language Processing, 20*(1), 14–22.
27. Dhanashri, D., and Dhonde, S. B. (2017). "Isolated word speech recognition system using deep neural networks". International Conference on Data Engineering and Communication Technology, Singapore, pp. 9–17.
28. Kaur, G., Srivastava, M., & Kumar, A. (2018). Integrated speaker and speech recognition for wheel chair movement using artificial intelligence. *Informatica, 42*, 587–594.
29. Trang, H., Loc, T.H., and Nam, H.B. (2014) "Proposed combination of PCA and MFCC feature extraction in speech recognition system". IEEE International Conference on Advanced Technologies for Communications, Hanoi, Vietnam, pp. 697–702.
30. Nikoskinen, T. (2015) "From neural network to deep neural network". Alto University School of Science, pp 1–27.
31. Gavat, I., and Militaru, D. (2015) "Deep learning in acoustic modeling for automatic speech recognition and understanding - an overview". IEEE International Conference on Speech Technology and Human-Computer Dialogue, Bucharest, Romania, pp. 1–8.
32. Sharmadha, S., Shivani, K., Shruthi, K., Bharathi, B., and Kavitha, S. (2020) "Automatic speech recognition using Deep Neural Network". International Conference on Soft Computing and Signal Processing, Hyderabad, India, pp. 353–361.

Dr. Gurpreet Kaur Dr. Gurpreet Kaur, born in 1982, is an Assistant Professor in the Department of Electronics and Communication Engineering at University Institute of Engineering and Technology, Panjab University, Chandigarh, India. She received her B.Tech (with Hons) in Electronics and Communication Engineering from Kurukshetra University, Haryana in 2004, M.E (with distinction) in Electronics and Communication from University Institute of Engineering and Technology, Panjab University, Chandigarh in 2007 and Ph.D. in Electronics Engineering from IKG Punjab Technical University, Jalandhar in 2018. She has more than 15 years of teaching experience. She has published more than 20 technical research papers in International and national Journals, Conferences and Seminars. Her current research interests are speech processing and neural networks.

Dr. Mohit Srivastava Dr. Mohit Srivastava, born in 1978, is a Professor in the Department of Electronics and Communication Engineering and Dean R&D at Chandigarh Engineering College, Landran, Mohali, Punjab, India. He received his B.Tech in Electronics and Communication Engineering from Magadh University, Bodh Gaya, M.Tech in Digital Electronics and Systems from K.N.I.T. Sultanpur and Ph.D. in Image processing & Remote Sensing from Indian Institute of Technology Roorkee in 2000, 2008 and 2013 respectively. He has more than 18 years of work experience at various environments includes Industries, educational and research centers. He has completed two IEDC (DST) funded project. He has published more than 25 technical research papers in International and national Journals, Conferences and Seminars. His current research interests are digital image and speech processing, remote sensing and their applications in Land Cover Mapping, and communication Systems.

Amod Kumar Dr. Amod Kumar did his B.E. (Hons.) in Electrical and Electronics Engineering from Birla Institute of Technology and Science, Pilani (Raj.); M.E. in Electronics from Punjab University, Chandigarh and Ph.D. in Biomedical Signal Processing from IIT Delhi. He has about 40 years of experience in Research and Development of different instruments in the area of Process Control, Environmental Monitoring, Biomedical Engineering and Prosthetics. He has worked as Chief Scientist at Central Scientific Instruments Organization (CSIO), Chandigarh which is a constituent laboratory of CSIR. Currently he is professor in Electronics department at NITTTR, Chandigarh. He has more than 70 publications in reputed national and international journals. He has worked at Technical University Berlin for one year on DAAD fellowship in 1987-88. He is also associated with Post Graduate Program of Academy of CSIR in the capacity of Professor. His areas of interest are Digital Signal Processing, Image Processing and Soft Computing.