ISBN: 978-81-19149-52-0

13. Isolated Spoken Word Transformation Using Feedforward Neural Network in Speaker Independent Speech Recognition

Satrughan Kumar Singh

PhD Research Scholar,
Department of Computer Science,
Central University of South Bihar,
Gaya, Bihar, India.

Anand Mohan

Principal,
Jinvani Management College,
Ara, Bhojpur, Bihar, India.

Muniyan Sundararajan

Professor,
Department of Mathematics & Computer Science,
Mizoram University,
Aizawl, Mizoram, India.

Abstract:

Advancements in growing technology and amidst the ongoing covid-pandemic period, the This paper presents a methodology for automated speech signal recognition of speaker independent robust isolated words. The use of voice command indeed contributes to a better integrated human-machine interface integration whereby one can give voice command which intelligent machine understand and obeys. It uses a feature set vector consisting of a combination of first-three formant of vocal speech signal tract. In this study, a feedforward neural network with back propagation (FFNN-BP) model based automatic speech-controlled system for speaker independent isolated spoken word recognition with mapping functions has been developed. This system is tested for the training and recognition of spoken commands and has given fairly accurate results with the precise operations of the automated speech recognition system for the given instruction.

Predicted prosodic parameters of the target isolated words independent of speakers are incorporated into neutral speech at multi-layered architecture to produce the resultant output. The whole process is evaluating quality assessment by subjective tests after incorporating inputted words or speech. The proposed model achieved an average recognition rate of 93.44% approximately. Finally, both subjective and objective tests reveal good and an increased accuracy with FFNN-BP model.

Keywords:

Word transformation; FFNN; zero crossing rate; linear predictive analysis; end point detection; multi-layer perceptron; MATLAB; speech recognition;

13.1 Introduction:

Machine learning is currently playing a vital role in the next generation of the computer world [1]. In the machine learning, the computer intelligence expert models are trained by known datasets of the domain context. The communication between human and computer occurs in both directions. This communication should have two important features of speech technology, speech recognition and speech synthesis. It is known that human uses emotions frequently to convey the intended message. Therefore, it is expected that the machine should be able to understand and generate desired emotions [2][11]. Most of the existing speech systems can generate only neutral style speech. In this situation, the transformation of emotion is applied to convert the neutral style speech to desired expressive style speech. Speech recognition can be classified into speaker-independent and speaker-dependent recognition. The speaker-dependent system must be trained for the specific users. This system is most practical choice in several applications and automations since it is very accurate in recognizing the user's voice or speech and may be used to provide a security level for access to some systems through spoken password. Although, speaker-independent system is user independent. Speech recognition can also be classified into single-word recognition (or isolated-word recognition) and continuous speech recognition. The continuous-speech recognition is going to be the ultimate voice or speech interface for the users' applications but nowadays there are some problems such as pause required between the words (while speaking rate restrictions on the users) to determine the end-points or boundaries of individuals words. Pronunciation of a given word which is affected by the words and punctuation around it making the recognition are quite difficult and inaccurate. The solution for this is widely accepted, highly accurate isolated-word recognition technique, which can be used to recognize each individual's words or short phrase by adding a short pause of few milliseconds between each utterance. Training facilitates a learning system for creating an acoustic training dataset [6]. From this, the meaning of entire phrase or sentence can be determined. An isolated-word recognition treats the word as the unit of recognition. That is, in the library are for the entire words rather than for phonemes or other constituents. This paper provides descriptive details for the implementation of neural network-based speaker-independent isolated-word recognition for a robot. Speech recognition in this application is done by considering different features such as STE, ZCR, maximum amplitude and LPA. Initially, few selected spoken words are stored as template speech. These features are then input to the FFNN for training and testing for the recognition of spoken commands. When a spoken command is reuttered, it is recognized by the neural networks.

13.2. Literature Review:

The speech recognition would be complicated enough even if every speaker pronounced every word in an identical manner each time, it was spoken. Successful speech recognition is dependent upon many factors including the application, the task, speaker-dependence,

size of vocabulary, speaker physical and emotional state, microphone quality, background noise, channel noise, details concerning language modeling, presence of confusable words in vocabulary, constraints imposed by the grammar and many others. The timing of utterance at recognition time will not be identical to that of during the training. These two utterances can have different duration. As a result, time-dependent features may fail to match because unknown and reference spoken words are out of phrase [12]. Despite the many sources of errors, it is possible to obtain an estimate of accuracy based on error rate measure. The solved problem on speech recognition is not simple due to both time dependence and information redundancy of the speech signal [3][9]. The detection of start and end points of an utterance is required in many of the speech recognition process to extract the speech data window (envelope of interest), which is noise-free. In the explicit methods, the analysis is done over the envelopes to estimate some features like short-time energy, zero crossing rate, etc. [4][5][7]. The function has been widely used for detection through short-time energy (STE). Once the detection function is obtained, it can be easily integrated in an adaptive threshold-based end-point detection system [10]. Optimum results were obtained in speech recognition when each envelope was divided into six frames and from each frame distinct parameters i.e., features were extracted [8]. These features were supplied as input to the neural network for the training. The new way of interaction with machines is speech recognition whereby one can give voice commands, which the machine understands and obeys.

13.3. Methodology:

In this study, a feedforward neural network with back propagation (FFNN-BP) model based automatic speech-controlled system for speaker independent isolated spoken word recognition with mapping functions has been developed. The classification is done using 5-layered FFNN architecture (see Figure 13.1). The classification step involves FFNN based mapping functions to classifying spoken words with multi-layered architectural framework of input, hidden and output layers by applying 10-fold matrix operation for training and testing datasets. These five layers (01 input layer, 03 hidden layers and 01 output layer) have 13, 20, 7, 20 and 13 neurons, respectively.

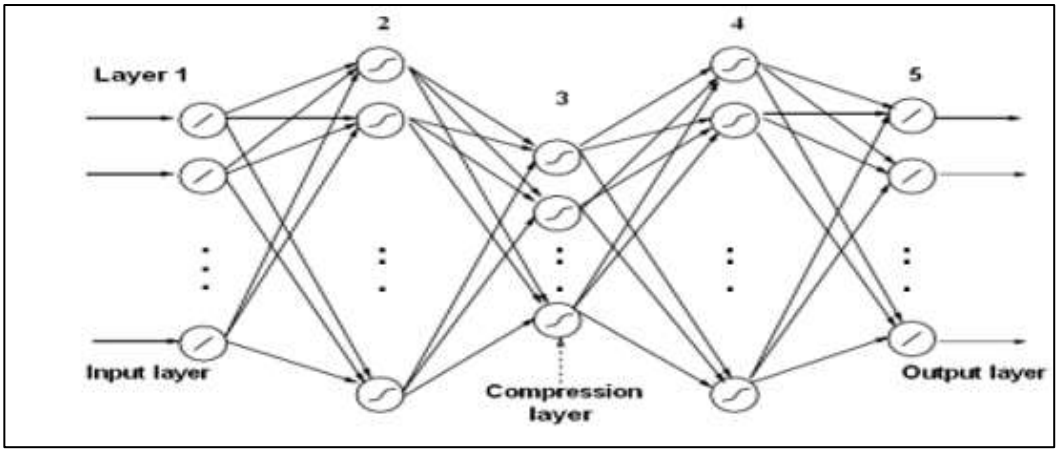

Figure 13.1: The Diagram 5 Layers Feed Forward Neural Network Model

In the proposed work, FFNN is used for recognition of eight different spoken commands. We have selected five sample utterances of six different speakers from the speech database collected from one male and one female speaker. For training & testing purpose, we have used 70 and 30 parallel utterances, respectively.

The multi-layered architecture incorporates the predicted prosodic parameters of target individual words into neutral speech, independent of the speakers, to produce the desired output. Subjective tests evaluate the perceptual quality of the transformed speech after inputted words or speech is included.

13.3.1. End Point Detection:

In speech signals the spacing between phonetic events will not be consistent. Words from different speakers of varying intensity flare up gradually to find the end-points of the energy signal. The start and end points exceed the minimum and maximum threshold values to identify the beginning of the word. The threshold value is obtained by calculating the normalized energy during the silence noise or background noise. The data in the speech lies between start point and end point i.e., speech envelope. This envelope is divided into small frames using a combination of STE, ZCR, maximum amplitude and LPA.

13.3.2. Sort Time Energy (STE):

STE is able to separate the speech segment that changes with every time slice from voiceless to voice. The energy value for the voiceless segment is much lower than voiced segment. The $\mathrm{E_n}$ allow us to distinguish between voice and unvoiced speech segments and locate approximately time of change between these two states. The mathematical calculation for STE is expressed as:

$$E_n = \sum_{m=-\infty}^{\infty} \mathrm{Y}^2(\mathrm{m}) * \mathrm{W}(\mathrm{n}-\mathrm{m})$$

where, $\mathrm{E_n}$ values allow us to distinguish between unvoiced and voice speech segments and locate approximately time of change between these two states, and $\mathrm{W}(\mathrm{n}-\mathrm{m})$ denotes finite length window.

13.3.3 Short Time Average Zero Crossing Rate (ZCR):

ZCR is involved in energy distribution with a strong correlation. A reasonable generalization is that if the short-time averages ZCR are high then the audio signal is not voiced, whereas if it is less, the audio signal is sounded. ZCR also help to characterize the fricatives or sibilants. It only requires a comparison of the signals of successive samples. The amplitude of the ZCR signal allows measurement of the number of times the signal crosses above zero by transition from positive direction to negative or vice versa. Here, rectangular window function employs to split the audio signal into temporal segments between the range of zero to positive direction or negative direction or vice versa. The mathematical calculation for zero crossing rate is expressed as:

$$Z = \sum_{i=1}^{\omega} \left(\frac{Sn(x_i) - Sn(x_{i-1})}{2} \right)$$

where $\mathrm{Sn}(\mathrm{x_n})$ denotes the sign of the $\mathrm{i^{th}}$ sample at the three possible values $+1, 0$ and -1.

13.3.4. Linear Predictive Coding (LPC) Analysis of Speech:

LPC is a most important measure and powerful pre-eminent technique of speech analysis for accurate and fast estimation of basic speech parameters. It is essentially a sequential formulation of problems related to speech wave modeling. One technique in linear predictive coding is auto-correlation method which is used in our application. A set of approximation coefficients is used to minimize the mean square error over a small segment of the waveform. A set of approximation coefficients is used to minimize the mean square error over a small segment $\mathrm{Y_n(m)}$ in the vicinity of sample n is uniformly 0 outside the interval $0 £ \mathrm{m} £ (\mathrm{N}-1)$, expressed as:

$$\mathrm{Y_n(m)} = (\mathrm{Y_n(m+n)} * \mathrm{w(m)})$$

Where, $\mathrm{w(m)}$ = finite length window in the interval $0 £ \mathrm{m} £ (\mathrm{N}-1)$.

The predictor function of $\mathrm{p^{th}}$ order is expressed as:

$$\mathrm{P(Z)} = \sum_{k=1}^{p} \mathring{a} * \mathrm{a_k} * \mathrm{Z^{-k}}$$

The corresponding prediction error, $\mathrm{e_n(m)}$ for the $\mathrm{p^{th}}$ order predictor will be non-zero on the interval $0 £ \mathrm{m} £ (\mathrm{N}-1) + \mathrm{p}$. The average prediction error of STA is expressed as:

$$\mathrm{\mathcal{E}_n} = \sum_{m=1}^{N+p-1} \mathrm{e_n^2(m)} = \sum_{m=1}^{N+p-1} \left\{ \mathrm{Y_n(m)} - \sum_{k=1}^{p} \mathrm{a_k} * \mathrm{Y_n(m-k)} \right\}^2$$

It can be shown that minimum mean square prediction error takes the form,

$$\mathrm{\mathcal{E}_n} = \mathrm{R_n(0)} - \sum_{k=0}^{p} \mathrm{a_k} * \mathrm{R_n(k)}$$

Where, $\mathrm{R_n(k)} = \sum_{\mathrm{m=0}}^{\mathrm{N-1-k}} (\mathrm{Y_n(m)} * \mathrm{Y_n(m+k)})$ is the auto-correlation value.

13.3.5 Mel Frequency Ceptral Coefficients (MFCC):

MFCC is an important measuring technique for time and frequency domain analysis of the audio signal processing which is divided into overlapping frames to calculate coefficients values. MFCC describes a Mel frequency cestrum that represents the short-time power spectrum of a signal. This spectrum is based on a linear cosine transform applied to a non-linear Mel frequency.

Let each frame have N number of samples and adjacent frames (multiplied by window) are separated by M number of samples, where M<N. The mathematical expression for the Hamming window is as follows:

$$\omega(n) = 0.5 - 0.4 * \cos(\frac{2\pi n}{N-1})$$

In MFCC, we use Fourier transform method to transform audio signal from time domain (TD) to the frequency domain (FD). Next, FD signal is converted to a matching frequency scale for appropriate perceptions and to find the final feature vector, $(p+1)+1+q+W+W$.

13.3.6. Feature Vector Classification and Recognition:

The recognition operation is divided into two phases, namely learning phase and classification phase. The weights in the neural network will be modified during the processing of learning phase. All weights are modified such that when a pattern is presented, the output is either the correct category or largest value.

The weight of the network remains constant in the classification phase. The input pattern will be transformed into layered network steps until it reaches the output layer. Finally, classification can be done by selecting the class associated with the output layer with the largest value or correct category. The classification phase is very fast phase than learning phase in term of fixed weights of the network and transformation from layer-to-layer iterations until correct and largest output value not achieved with back propagation architectures. The feature vectors consist of two elements, formant frequencies and mean ZCR of the file. The formant frequency is categorized with three elements, represented as F,

$$F = \{ff1, ff2, ff3, Z\}$$

13.4 Results & Discussions:

For implementing proposed FFNN-BP based soft computing system, we have considered MATLAB programming language. In this system, FFNN-BP is used for training and classification of isolated spoken words of eight spoken commands by the different speakers. In this proposed system, we used 5 sample utterances of 6 different users for each of 8 commands (240 speech sample files).

Optimum results were obtained when each speech envelope was divided into six frames and from each frame 7 parameters were extracted (4 pole LPC, ZCR, total energy and maximum amplitude features). The sets of total 42 parameters were given to neural network for 6 different users.

The neural network recognizes the commands and input features extracted from speech envelope. The model is tested with data validation for audio commands and has given fairly accurate results with the precise operations of the automated speech recognition system for the given instruction.

The entire process is evaluated by subjective tests after incorporating inputted words or speech. It uses confidence measures with mapping functions to detect and classify the spoken words by the subjective and objective evaluations.

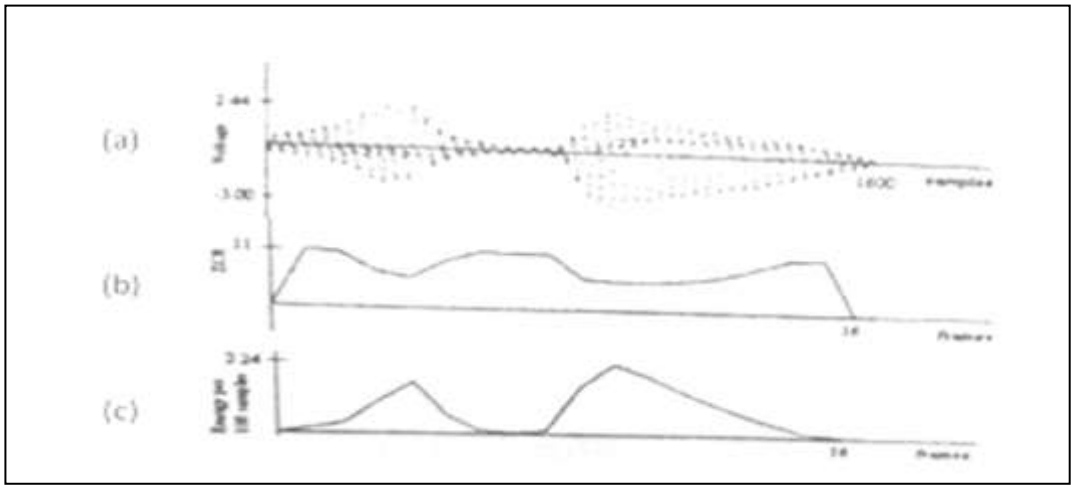

Figure 13.2: (a) Original Signal (b) ZCR of the Signal (c) Energy of the Signal

Figure 13.2 shows plots of number of samples of original signal versus amplitude, ZCR and energy of the signal. It can be seen that the energy contained in the silence period is small as compared to that in the utterance period. Also, the energy depends on the intensity of the utterance.

Figure 13.3 shows the training and testing plots for different features.

Figure 13.3(a) indicates the variation in the value of the first formant frequency ff1 for each of the four words uttered by the six speakers during the training phase while Figure 13.3(b) indicates the same for the other set of six speakers used for the testing phase.

Similarly, Figure 13.3(c) and Figure 13.3(d) shows the variations of the second formant frequency ff2. Figure 13.3(e) and Figure 13.3(f) shows the variations of the third formant frequency ff3, and Figure 13.3(g) and Figure 13.3(h) shows the variations of the mean ZCR during the training and testing phases respectively.

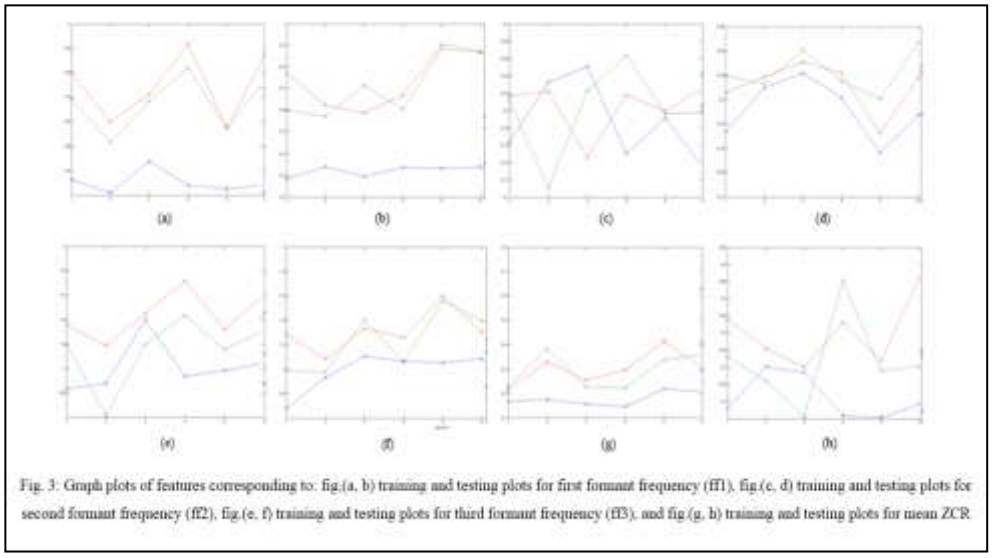

Fig. 3. Graph plots of features corresponding to: fig.(a, b) training and testing plots for first formant frequency (ff1), fig.(c, d) training and testing plots for second formant frequency (ff2), fig.(e, f) training and testing plots for third formant frequency (ff3), and fig.(g, h) training and testing plots for mean ZCR

Figure 13.3: Graph plots of feature corresponding to: Figure (a, b) training and testing plot for first format frequency (ff1), Figure (c, d) training and testing laws for second format frequency (ff2), Finger (e, f) training and testing floor of current format frequency (ff3), and Figure (g, h) training and testing plot for mean ZCR.

Thus, energy has been set as a major criterion for the end-point detection. Zero crossing rate of an unvoiced signal is more compared with the zero-crossing rate of voiced signal. It is also observed that the zero crossing rates for different words are different. From the various signal interpolation with traversing data plotting, it has been deducted that the speech envelopes differ with various words and hence the number of samples vary from word-to-word, time-to-time and person-to-person. The overall average accuracy of the proposed model is achieved by 93.44% approximately. Finally, both subjective and objective tests reveal good and an increased accuracy with FFNN-BP based mapping methods.

13.5 Conclusions and Future scope:

Automatic speech recognition is a challenging task for spoken words of different speakers because human speech signals are highly variable due to different aspects such as, speaker attributes, spoken styles, uncertain noises, and many. The decisions for classifying infant cries are quite difficult. In this study, the performance of FFNN-BP model is accessed that classified and recognize the spoken words or commands of the different vocal tracts by applying machine learning algorithms.

In this system, FFNN-BP based automatic speech recognition tool provides fairly accurate results to understand the classification of spoken commands of different speakers on different vocal tract and emotions for isolated words transformation.

We have looked at different papers and noticed that different authors have also used different methodologies and provided different accuracy.

So, we can say that methodology and accuracy both are complementary to each other and if methodology is changed then accuracy will also change and the same conclusion we have obtained from different paper of study. In this work, the mapping function map the associations between spectral parameters of neutral and target isolated words by subjective and objective tests after incorporating inputted spoken words.

The system is tested for recognition and has given fairly accurate result with average recognition rate of 93.44% approximately. Both subjective and objective tests reveal effective results through FFNN-BP based mapping methods. Prosodic and spectral parameters contain complementary emotional information. Thus, the proposed model describes an automated system which is suited to an effective performance for speech signal recognition of isolated words independent of different speakers on different vocal tract by training and testing the spoken word or commands datasets. In the future work, both spectral and prosodic mappings will be performed together to improve the more performance of subjective and objective test evaluations by training and testing large datasets. Many other applications such as voice-controlled wheel chair, voice-interactive enquiry system for railway, bank and insurance sectors, etc. can be developed using similar approach.

13.6 References:

1. S.K. Singh and J. Yadav, "Machine Learning & Image Processing for hand written digits and alphabets recognition from document image through MATLAB simulation", IOP Conf. Series: Materials Science and Engineering, Vol. 1084(2021)012021:1-8, pp. 1-8, 2021. DOI:10.1088/1757-899X/1084/1/012021
2. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction", Signal Processing Magazine, IEEE, vol. 18, pp. 32–80, 2001.
3. J. Psutka, "A Limited Vocabulary Speech Recognition System", In: Proceedings of the Third International Conference on Artificial and Information Control Systems of Robots, Czechoslovakia, pp. 301-304, 1984.
4. S. Furvi, "Speaker-Independent Isolated-Word Recognition Using Dynamic Features of Speech Spetrum", IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-34, pp. 53-59, 1986.
5. L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Englewood Cliffs, NJ: Prentice-Hall, 1978.
6. S.K. Singh, M. Sundararajan and J. Yadav, "An automatic infant cry speech recognition using artificial neural network", In: 2nd International Conference on Multidisciplinary Research and Studies, Indirapuram Institute of Higher Studies, Ghaziabad, Uttar Pradesh, India, 20th August 2022.
7. T. Parson, "Voice and Speech Processing", McGraw-Hill Book Company, 1987.
8. M.D. Tom and M.F. Tenorio, "Experiments with the Spatio-temporal Pattern Recognition Approach and Dynamic Time Warping Approach to Word Recognition", IEEE Trans. On Acoustic, Speech, and Signal Processing, pp. 445-448, 1989.
9. M. Schrder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," In: Proceedings of the Eurospeech, pp. 87 –90, 2001.
10. N.P. Jawarker and A. Vasif, "Smart Voice Operated (Speaker Independent) Automated Telephone Call Establishment Machine Using Neural Network", In: Proceedings of the

National Conference on Applications of Neural Networks and Fuzzy Systems, Annamalai University, Feb 2002.
11. G.E. Pelton, "Voice Processing", McGraw Hill, 1993.
12. M.J. Navarro, B.A. Moreno and S.E. Lleida, "An Improved Speech End-point Detection System in Noisy Environments by Means of Third Order Spectra", IEEE Trans. Signal Processing, Vol. 6, No. 9, pp. 224-226, 1999.