Vowel Classification based on LPC and ANN

R. B. Shinde
College of Computer Science and Information Technology, Latur, Maharashtra.

Dr. V. P. Pawar
Associate Professor in Computer Science Dept of SRTM University, Nanded, Maharashtra.

ABSTRACT
The vowel sounds are perhaps the most interesting class of sound in English. Their importance to the classification and representation of written text is very low; however, most practical speech recognition systems rely heavily on vowel recognition to achieve high performance. In this paper we propose a technique for the vowel classification using Linear Prediction Coefficient with combination of statistical approach and Artificial Neural Network. The proposed technique achieves the 98.7% accuracy rate for vowel classification.

General Terms
ANN (Artificial Neural Network), Discrete Cosine Transform (DCT), Fast Fourier Transform (FFT), Linear Predictor Coefficients (LPC),

Keywords
Vowel Classification System (VCS),

1. INTRODUCTION
In speaking, vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air caused by the vibration of the vocal cords. The way in which the cross-sectional area varies along tract determine the response frequencies of the tract i.e. the formants and thereby the sound that is produced. The vowel sound produced is determined primarily by the position of the tongue, but the position of the jaw, lips and to a small extant the velum, also influence the resulting sound. [1] Figure 1 shows the schematic view of the human vocal mechanism

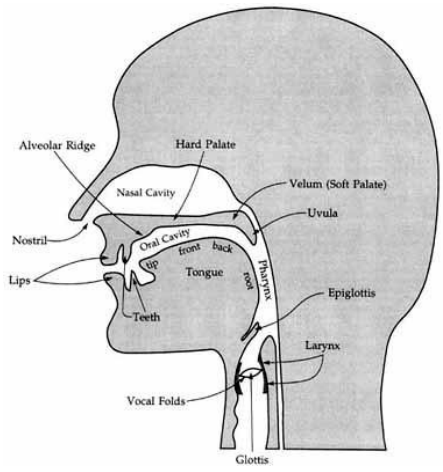

Fig 1:- Schematic View of the Human Vocal Mechanism

Vowels are most important class of sound in most of the world languages. The duration of the vowel is longer than consonants and is most significant. As the vowels are easily & reliably recognized therefore they are used to recognize speech by the human beings. The International Phonetic Association maintains the International Phonetic Alphabet, or IPA, a standard set of characters to describe the sounds of human speech. The IPA separates phonemes into consonants and vowels, as well as a few other types of sounds that are not present in English, and then arranges each group according to the way the phonemes are constructed by the speaker. Vowels are described in terms of three parameters: the openness of the speaker's mouth when speaking the vowel, the position in the mouth where the sound is generated, and whether the mouth is rounded when producing the sound. See Fig 2.

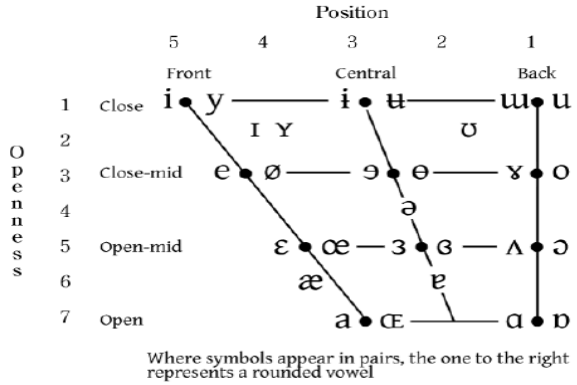

Fig 2:- The IPA describes vowels in terms of three parameters: openness, position, and roundness [2].

The paper is divided into five sections, section 1 for Introduction, section 2 deals with the Data base generation, section 3 deals with the vowel classification system, section 4 covers the result and section 5 gives the conclusion

2. DATA BASE GENERATION
The speaker were seating on the chair with relax position wearing the headphones for recording the vowels. Here we use headphones for recording so that for each speaker equal distance between mouth & speaker is kept automatically. The sampling frequency for all recording was 11025 Hz. The speech data is collected with the help of sound recording software. Sound files are recorded in the PCM format and saved with the extension .wav. For this experiment we took the samples of 5 persons. 5 samples of each vowel i.e. 'A', 'E', 'I', 'O', 'U' basic vowels. Signals have been sampled to 16 kHz with an analysis LPC, takes all 200ms in Hamming windows of 10 ms giving each 10 LPC coefficients and the corresponding residual energy (13th parameter). Saved files are labeled properly and these files are stored in memory for further processing. This work was achieved on windows microcomputer 2.13 GHz with 3.00GB of RAM, developed by the C++ builder and Matlab 2009.

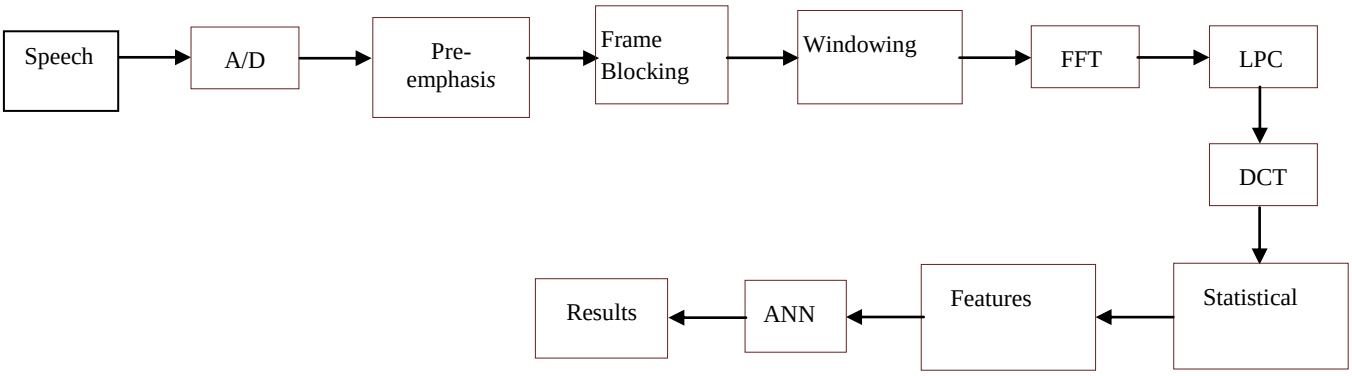

Fig. 3: Block structure for the VCS

3. BUILDING VOWEL CLASSIFICATION SYSTEM (VCS)

During the first step i.e. speech acquisition, speech samples are obtained from the speaker in real time and stored in memory for preprocessing. Building VCS comprises mainly two parts first part deals with the feature extraction and second part deals with the classifier. Following are the steps for the feature extraction & classifier and these are presented by the Fig. 3

3.1 Feature Extraction

Feature extraction is the most important phase in the speech processing. Vowel recognition is the process of recognizing which vowel is uttered based on unique characteristics contained in speech waves. Many contributions give a great interest in artificial neural network field because it has demonstrated that connectionist architectures are capable of capturing some critical aspects of the dynamic nature of speech, can achieve superior recognition performance for difficult but small phonemic discrimination tasks such as discrimination of the voiced consonants /B/, /D/ and /G/ [3].

There are many techniques used to parametrically represent a voice signal for Vowel classification tasks. These techniques include Linear Predictor Coefficients (LPC), Auditory Spectrum-Based Speech Feature (ASSF), and the Mel-Frequency Cepstrum Coefficients (MFCC). [8] The LPC technique is used in this paper. Following are the involved in the feature extraction.

1. *Analog to digital convertor:* for converting the analog speech to digital signal we use the headphones & sound recording software.
2. *Pre-emphasis wave:* The digitized speech signal, s (n), is put through a low-order digital system to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. Pre-emphasis network is the fixed first-order-system.

$$H(z) = 1 - \tilde{a}z^{-1}, \quad 0.9 \le a \le 1.0 \qquad (0.1)$$

In this case, the output of the pre-emphasis network, s (n), is related to the input to the network, s (n) by the difference equation

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \qquad (1.2)$$

The most common value for a =0.95. A simple example of a first-order adaptive pre-emphasizer is the transfer function.

$$H(z) = 1 - \tilde{a}_n z^{-1}, \qquad (1.3)$$

Where $\tilde{a}_n$ changes with time (n) according to the chosen adaption criterion. [10]

3. *Sampling:* Sampling is a process of converting a continuous-time signal into a discrete-time signal. It is convenient to represent the sampling operation by a fictitious switch. The switch closes for a very short interval of time T, during which the signal presents at the output. The time interval between successive samples is T seconds and the sampling frequency if given by

$$f = \frac{1}{T} \text{ Hz.} \qquad (1.4)$$

4. *Framing:* The process of segmenting the speech samples obtained from an ADC into a small frame with the length within the range of 20to 10 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M .To avoiding the frame overlapping problem the frame is shifted every 10 samples. The used values for N & M are 200ms & 10ms when the sampling rate of speech is 11025 Hz.
5. *Windowing:* Next process is to apply window to each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. A Hamming window is used for autocorrelation method in LPC. Hamming window has the form as given below.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \qquad (1.5)$$

$$0 \le n \le N-1.$$

6. *Fast Fourier Transform:* To convert each frame of N samples from time domain into frequency domain FFT is being used. FFT is used to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain. The equation (1.6) obtains the value of FFT.

$$S(n) = \sum_{j=1}^{N} s(j)\omega_N^{(j-1)(k-1)} \qquad (1.6)$$

$$s(j) = (1/N)\sum_{k=1}^{N} S(k)\omega_N^{-(j-1)(k-1)} \qquad (1.7)$$

Where

$\omega_N = e^{(-2\pi i)/N}$ is an Nth root of unity.

7. *Linear predicted Co-efficient:* LPC determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. It finds the

coefficients of a nth-order linear predictor that predicts the current value of the real-valued time series $\tilde{s}(n),$ based on past samples.[17]

$$\tilde{s}(n) = -A(2) * X(n-1) - A(3) * X(n-2) - .... - A(N+1) * X(n-N) \quad (1.8)$$

n is the order of the prediction filter polynomial, a = [1 a(2) ... a(p+1)]. If n is unspecified, LPC uses as a default n = length(x)-1. If x is a matrix containing a separate signal in each column, LPC returns a model estimate for each column in the rows of matrix and a column vector of prediction error variances. The length of n must be less than or equal to the length of x.

8. *Discrete Cosine Transform:* DCT can be used to achieve the coefficients. DCT reconstruct a sequence very accurately from only a few DCT coefficients, a useful property for applications requiring data reduction. DCT returns the discrete cosine transform of X. The vector Y is the same size as X and contains the discrete cosine transform coefficients[16]

$$y(k) = wk \sum_{n=1}^{n} x(n) \cos \frac{\pi(2n-1)(k-1)}{2N}$$

k=1,., N. (1.9)

Where $wk = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \frac{\sqrt{2}}{N} & 2 \le k \le N \end{cases}$

9. *Applying Statistical parameters:* For this proposed work we use a simple statistical parameter Standard deviation. After applying the DCT on the speech signal a matrix is produced and generally it's very difficult to operate on such a large data. So we reduce that data by using standard deviation & we extract the 10 features for further work. Now we have total 75 samples from the 5 different speakers. In this way we get total 375 features from our 5 speakers.

3.2 Artificial Neural network

Artificial Neural Networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous system. Commonly, neural networks are adjusted, or trained, so that a particular input leads to a specific target output. In this paper we use a two-layer feed-forward network, with sigmoid hidden and output neurons, can classify vectors arbitrarily well, given enough neurons in its hidden layer. The network will be trained with scaled conjugate gradient back propagation. In this algorithm, input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with a specific output vectors, or classify that vectors in an approximate way. The architecture of network used for a two-layer feed-forward network algorithm has been given in Fig 3. An elementary neuron with X1, X2 … inputs has been shown. Each input is weighted with an appropriate value wij.

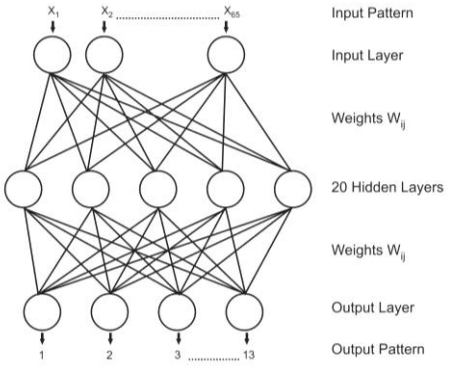

Fig. 3: A two-layer feed-forward network with 20 hidden neuron.

3.3 Confusion matrix

We have 53 examples vowels in learning and 11 examples for test. When we observe the confusion matrices for both training and test phases, we conclude that the principal confusion is caused by the I vowel phoneme for example in table 3, it take 3 examples of vowel I but one sample get misclassified so test result is 90.9% & error rate is 9.1% . Table 2 shows the 100% recognition rate in learning phase.

Table 2. Confusion matrix for learning examples

	A	**E**	**I**	**O**	**U**	
A	11 20.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
E	0 0.0%	8 15.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
I	0 0.0%	0 0.0%	12 22.6%	0 0.0%	0 0.0%	100% 0.0%
O	0 0.0%	0 0.0%	0 0.0%	14 26.4%	0 0.0%	100% 0.0%
U	0 0.0%	0 0.0%	0 0.0%	0 0.0%	8 15.1%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Table 3. Confusion matrix for test examples

	A	**E**	**I**	**O**	**U**	
A	2 18.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
E	0 0.0%	3 27.3%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
I	0 0.0%	1 9.1%	2 18.2%	0 0.0%	0 0.0%	66.7% 33.3%
O	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
U	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 27.3%	100% 0.0%
	100% 0.0%	75.0% 25.0%	NaN% NaN%	100% 0.0%	100% 0.0%	90.9% 9.1%

Hypotheses:-

Each vowel sound in English is located in a distinct region of some high-dimensional phonological space that maximizes distances between each pair of distinct vowels [4]. Therefore it is possible to programmatically segment this space to classify vowel sounds.

- The classification parameters used to describe vowels in the International Phonetic Alphabet approximate the phonological space in which vowel sounds are naturally described.
- The first thirteen mel-frequency cepstral coefficients of a vowel sound describe its unique properties sufficiently to differentiate it from other vowels [9].
- Training the ANN with data from multiple speakers will force it to correlate speaker-independent common properties of vowel sounds. Thus it will be able to classify vowels spoken by individuals whose voices were not included in the training data set.

4. RESULTS

The vowel classification results were obtained using the generated database. 10 features from each vowel have been extracted using LPC. Total 75 samples are used for pattern recognition.75 samples are classified into 5 vowel classes. In pattern recognition problems a neural network is used to classify inputs into a set of target categories. The proposed features have been tested on an Artificial Neural Network Using a MATLAB tool. The Neural Network Pattern Recognition Tool will help to select data, create and train a network, and evaluate its performance using mean square error and confusion matrices. The result of the Speaker recognition is shown below Fig 3(a), Fig 3(b) respectively in the form of confusion matrix and mean square error.

Table 3. Confusion matrix for train examples

A	15 20.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
E	0 0.0%	14 18.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
I	0 0.0%	1 1.3%	15 20.0%	0 0.0%	0 0.0%	93.8% 6.3%
O	0 0.0%	0 0.0%	0 0.0%	15 20.0%	0 0.0%	100% 0.0%
U	0 0.0%	0 0.0%	0 0.0%	0 0.0%	15 20.0%	100% 0.0%
	100% 0.0%	93.3% 6.7%	100% 0.0%	100% 0.0%	100% 0.0%	98.7% 1.3%

Fig. 3(a): Confusion Matrix displaying 98.7% recognition rate

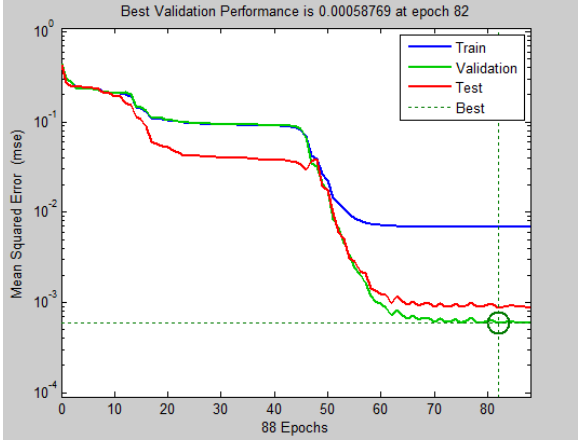

Fig. 3(b): Mean Square Error (MSE) displaying best validation performance at epoch 102

From the confusion matrix it is cleared that the 65 samples from the 13 speaker each is correctly classified into 13 classes only one sample is get misclassified & recognition rate is obtained 98.7% & 1.3% error rate. Mean Squared Error is the average squared difference between outputs and targets. Lower values are better. Zero means no error.

5. CONCLUSION

The maximum average accuracy achieved for any network was 91.5%. This network had a single layer of 20 internal nodes and was trained to distinguish between five unique vowels. The network was trained for a total of 88 iterations using the training data. The vowel the network recognized with the greatest reliability was 'A', 'E', 'O', 'U' at 100%. The lowest accuracy for a single vowel with this network was 'I' at 93.3%. Our success rates are summarized in figure 8 below.

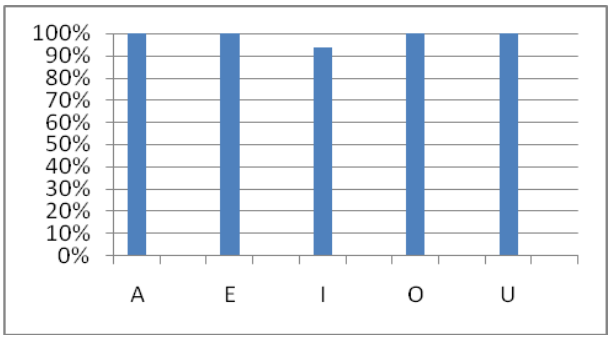

Figure 4: Recognition success rates with a feed-forward ANN.s

The average correct recognition rate we have obtained was 93.23% (98.7% the best and 87.77% the worst), and the largest wrongly recognized vowel percentage was 6.3% for the vowel 'I'. This result leads us to conclude that the described algorithm can be implemented to systems for Automatic recognition of vowels in continuous speech of English language. This algorithm can be easily applied to other languages too

6. REFERENCES

[1] International Phonetic Association, Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece. International Phonetic Alphabet chart, 1996.

[2] A. Waibel, H. Sawai, and K. Shikano, "Consonant Recognition by Modular construction of Large Phonemic Time Delay Neural Network", IEEE International Conference on Acoustics, Speech, and Signal Processing , pp. 107-110, April 1989.

[3] Gopalakrishna Anumanchipalli, Rahul Chitturi, "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems"

[4] Singh, S. P., et al Building Large Vocabulary Speech Recognition Systems for Indian Languages, International Conference on Natural Language -Processing, 1:245-254, 2004.

[5] Evgeniy Gabrilovich, Alberto D. Berstin: "Speaker recognition: using a vector quantization approach for robust text-independent speaker identification", Technical report DSPG-95-9-001", September 1995.

[6] Tridibesh Dutta, "Text dependent speaker identification based on spectrograms", Proceedings of Image and vision computing, New Zealand 2007.

[7] D. A. Reynolds, "An overview of automatic speaker recognition technology," Proc. IEEE Int. Conf. Acoustic., Speech Signal Process. (ICASSP'02), 2002, pp. IV-4072–IV-4075.

[8] Jamel Price and Ali Eydgahi, "Design of Matlab®-Based Automatic Speaker Recognition Systems," 9th International Conference on Engineering Education T4J-1, July 23 – 28, 2006.

[9] "Isolated Word, Speech Recognition using Dynamic Time Warping." Dynamic Time Warping. 14 June 2005.

[10] Jiehua Dai Zhengzhe Wei, "Study and Implementation of Feature Extraction and Comparison In Voice Recognition"

[11] Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C. Mehrotra, "Marathi Isolated Word Recognition System using MFCC and DTW Features", Proc. of Int. Conf. on Advances in Computer Science 2010

[12] Qi li, frank k. Soong, and olivier siohan, "A High-Performance Auditory Feature For Robust Speech Recognition"

[13] Transform Dr. H B Kekre1, Vaishali Kulkarni, "Speaker Identification using Row Mean of DCT and Walsh Hadamard" International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 3 Mar-2011

[14] "Digital Signal Processing" By-P.Ramesh Babu Scitech Publications (India) PVT, LTD.

[15] "Fundamental of Speech Recognition" By-Lawrence Rabiner , Biing-Hwang Juang, Published by Pearson Education (Singapore) Pte. Ltd. Indian Branch.

7. AUTHORS PROFILE

Mrs. R. B. Shinde received the M.Sc. (CS) degree from Dr. B. A. M. University, Aurangabad, in the year 2001. She is currently working as lecturer in the College of Computer Science and Information Technology, Latur, Maharashtra. She is leading to Ph. D degree in S.R.T.M. University, Nanded.

Dr. Vrushsen V. Pawar received MS, Ph.D. (Computer) Degree from Dept .CS & IT, Dr. B. A. M. University & PDF from ES, University of Cambridge, UK. Also Received MCA (SMU), MBA (VMU) degrees respectively. He has received prestigious fellowship from DST, UGRF (UGC), Sakaal foundation, ES London, ABC (USA) etc. He has published 90 and more research papers in reputed national international Journals & conferences. He has recognized Ph. D Guide from University of Pune, S. R. T. M. University & Singhaniya University (India). He is senior IEEE member and other reputed society member. Currently working as Associate Professor in CS Dept of SRTMU, Nanded.