International Journal on Soft Computing (IJSC) Vol. 5, No. 3, November 2014

BROAD PHONEME CLASSIFICATION USING SIGNAL BASED FEATURES

Deekshitha G[1] and Leena Mary[2]

[1,2]Advanced Digital Signal Processing Research Laboratory, Department of Electronics and Communication, Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India

ABSTRACT

Speech is the most efficient and popular means of human communication Speech is produced as a sequence of phonemes. Phoneme recognition is the first step performed by automatic speech recognition system. The state-of-the-art recognizers use mel-frequency cepstral coefficients (MFCC) features derived through short time analysis, for which the recognition accuracy is limited. Instead of this, here broad phoneme classification is achieved using features derived directly from the speech at the signal level itself. Broad phoneme classes include vowels, nasals, fricatives, stops, approximants and silence. The features identified useful for broad phoneme classification are voiced/unvoiced decision, zero crossing rate (ZCR), short time energy, most dominant frequency, energy in most dominant frequency, spectral flatness measure and first three formants. Features derived from short time frames of training speech are used to train a multilayer feedforward neural network based classifier with manually marked class label as output and classification accuracy is then tested. Later this broad phoneme classifier is used for broad syllable structure prediction which is useful for applications such as automatic speech recognition and automatic language identification.

KEYWORDS

Automatic Speech Recognition, Broad phoneme classes, Neural Network Classifier, Phoneme, Syllable, Signal level features,

1. INTRODUCTION

In terms of human communication, speech is the most important and efficient mode of communication even in today's multimedia society. So we want automatic speech recognition systems to be capable of recognizing fluent conversational speech from any random speaker. Speech recognition (SR) is the translation of spoken words into text. In the ASR (Automatic speech recognition) system, the first step is feature extraction, where the sampled speech signal is parameterized. The goal is to extract a number of parameters/features from the signal that contains maximum information relevant for the classification. The most popular spectral based parameter used in state-of-art ASR is the Mel Frequency Cepstral Coefficients (MFCC) [1][2]. The main demerits of MFCC include its complex calculation and limited recognition rate. The poor phone recognition accuracy in conventional ASR is later compensated by the language models at sub word and word levels to get reasonable word error rate [3].

Speech is composed of basic sound units known as phonemes [1][2]. The waveform representation of each phoneme is characterized by a small set of distinctive features, where a distinctive feature is a minimal unit which distinguishes between two maximally close but linguistically distinct speech sounds. These acoustic features should not be affected by different

DOI: 10.5121/ijsc.2014.5301

vocal tract sizes and shapes of speakers and the changes in voice quality. In this work, we have attempted broad phoneme classification using signal level features. Broad phoneme classes include vowels, nasals, plosives, fricatives, approximants and silence [1][2]. Each of these classes has some discriminant features so that they can be easily classified. For example vowels can be easily categorized with its higher amplitude, whereas fricatives with their high zero crossing rate. So for identifying such characteristics an analysis study was conducted and results were summarized. From this study, we have come up with a set of feature vectors capable of performing broad phoneme classification.

The paper is organized as follows: Section 2 gives the system overview by explaining the database, relevant features extraction and classifier. In Section 3, details of experiments are described with evaluation results. Finally in Section 4, the paper is wrapped up with a summary and scope for future work.

2. SYSTEM OVERVIEW

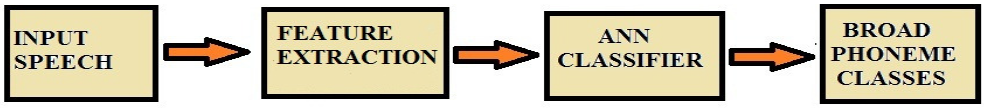

Figure 1. Broad phoneme classification

A classifier system for acoustic-phonetic analysis of continuous speech is being developed to serve as part of an automatic speech recognition system. The system accepts the speech wave-form as an input and produces a string of broad phoneme-like units as output. As shown in Figure 1, the input speech is parameterized in to normalized feature set [4][5], and then this feature is applied to the ANN [6] classifier in order to obtain the broad phoneme class labels.

Table 1. Broad phoneme classes considered in the work.

Sl No	Broad Phoneme Class	Symbols used to represent the class	IPA symbols of phonemes within each class	Corresponding Malayalam characters
1	*Vowels*	V	a i u e o a: i: u: e: o:	അ ഇ ഉ എ ഒ ആ ഈ ഊ ഏ ഓ
2	*Nasal*	N	m n̪ n ɳ ɲ ŋ	മ ന ന ണ ഞ ങ
3	*Stop*	P	b d̪ ɖ dʒ g bʰ d̪ʰ ɖʰ dʒʰ gʰ p t̪ ʈ tʃ k pʰ t̪ʰ ʈʰ tʃʰ kʰ t	ബ ദ ഡ ജ ഗ ഭ ധ ഢ ഝ ഘ പ ത ട ച ക ഫ ഥ ഠ ഛ ഖ ഠ
4	*Fricative*	F	f s ʂ ʃ h	ഫ സ ഷ ശ ഹ
5	*Approximant*	A	ʋ ɻ j l ɭ r ɾ	വ ഴ യ ല ള ര റ
6	*silence*	S		

The broad phoneme classes in Malayalam language that are considered in this work, their symbols and corresponding International Phonetic Alphabet (IPA) [1] symbols are listed in Table 1. The broad classifications yield the following categories: vowels, nasals [7], fricatives, plosives, approximants [5], and silence leading to six broad categories. For classifying the input sound units into one among the six classes, specific features are needed to distinguish them. So a study was conducted first by analyzing pitch contour, energy contour, formant plot, frequency spectrum etc.

2.1. Data Collection and Transcription

First phase of the work was the data collection and its manual transcription which was much time consuming. For speech data, Malayalam read speech was downloaded from All India Radio (AIR), and speech was also recorded in the lab environment. Figure 2 illustrates the manual labeling of speech signal using wavesurfer.

2.2. Feature Extraction

A preliminary analysis study was conducted to identify discriminative features helpful for broad phoneme classification. Various features [4][5][6] explored include the following: voicing information, short time energy, Zero Crossing Rate (ZCR), Most Dominant Frequency (MDF), magnitude at the MDF, spectral flatness measure, formant frequencies, difference and ratios of formant frequencies, bandwidth of formant frequencies etc [1][2]. Among them we selected nine features for designing the broad phoneme classifier. It includes (1) voicing decision, (2) ZCR, (3) short time energy, (4) most dominant frequency, (5) magnitude at MDF, (6) spectral flatness measure, and (7) first three formants. Brief descriptions about these selected features are explained below.

2.2.1. Voiced/Unvoiced Decision

Voicing information helps to determine whether the frame is voiced or unvoiced. Here we extracted pitch information using auto correlation function. Autocorrelation sequence is symmetric with respect to zero lag. The pitch period information is more pronounced in the autocorrelation sequence compared to speech and pitch period can be computed easily by a simple peak picking algorithm [6].

2.2.2. Zero Crossing Rate (ZCR)

The high and low values of ZCR correspond to unvoiced and voiced speech respectively as shown in Figure 3. In a speech signal *S (n)*, a zero-crossing occurs when the waveform crosses the time axis or changes algebraic sign [1].

$$ZCR(n) = \sum_{m=-\infty}^{\infty} 0.5 \,|\, \mathrm{sgn}(S(m)) - \mathrm{sgn}(S(m-1)) \,|\, w(n-m)$$

$$\mathrm{sgn}(S(n)) = \begin{cases} 1, \; for S(n) \geq 0 \\ -1, \; otherwise \end{cases}$$

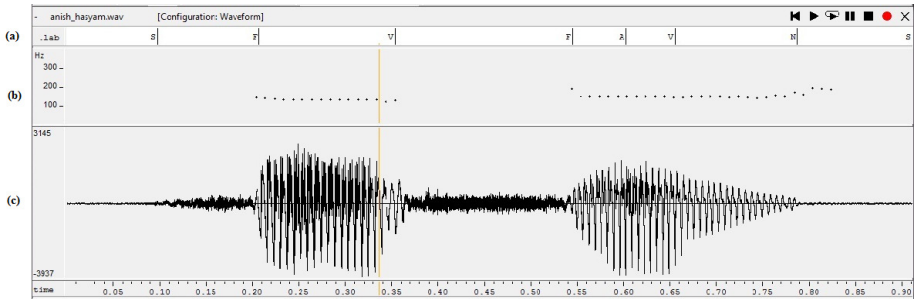

Figure 2: (a) Manual labelling to broad phonemes, (b) pitch contour and (c) corresponding speech waveform.

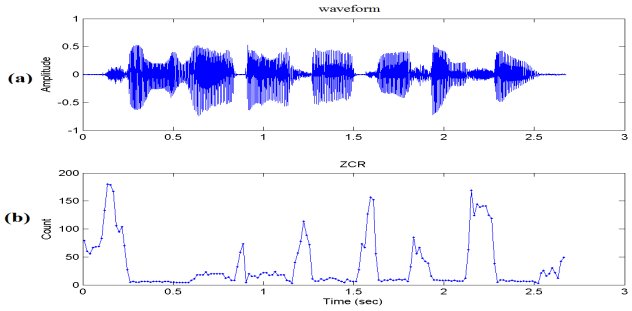

Figure 3. (a) Speech waveform and (b) ZCR.

2.2.3. Short Time Energy (STE)

Energy can be used to segment speech into smaller phonetic units in ASR systems. Short Time Energy is a squaring or absolute magnitude operation. Energy emphasizes high amplitudes and is simpler to calculate. The large variation in amplitudes of voiced and unvoiced speech, as well as smaller variations between phonemes with different manners of articulation, permit segmentations based on energy [1][2].

$$E(n) = \sum_{m=-\infty}^{\infty} S(n)^2 w(n-m)$$

2.2.4. Most Dominant Frequency (MDF)

The most dominant frequency is the frequency of sinusoidal component with highest amplitude. If the signal is highly periodic and more or less sinusoidal, the dominant frequency will in most cases be related to the rate of the signal. The more a signal looks like a sine wave; the better dominant frequency analysis will be able to reflect the periodicity of the signal.

2.2.5. Spectral Flatness Measure (SF)

Spectral flatness is a measure used to characterize an audio spectrum. Spectral flatness is typically measured in decibels, and provides a way to quantify how tone-like a sound is, as opposed to being noise-like. A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands i.e., similar to white noise. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands i.e., a mixture of sine waves.

The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum.

$$SpectralFlatness = \frac{GM}{AM}$$

2.2.6. Formant Frequencies

Formants are the spectral peaks of the sound spectrum of the voice. It is often measured as amplitude peaks in the frequency spectrum of a sound. The behaviour of the first 3 formants is of crucial importance. Figure 4 justifies the selection of formant frequency [7] as a feature since it classifies vowels and nasals.

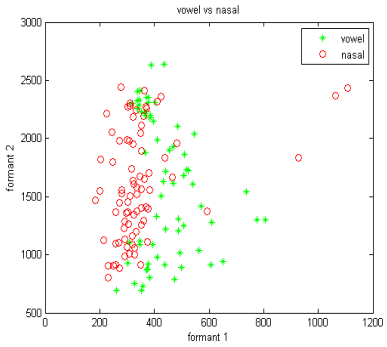

Figure 4.Vowel-nasal classification using first two formants

An optimum set of features has to be found, which describes the classes with a few, but very powerful parameters whereby showing optimum class discrimination abilities [6]. Thus, the feature extraction process can also be seen as a data reduction process. For getting the features from the input speech signal, first the signal is normalized and then it is segmented into frames of size 20ms with a 10ms overlap. Then for each of this frame, the features are calculated. The choice of the features is done after conducting a study which is summarized in Table 2. Figure 5 shows the plot of selected features for a particular word.

Table 2. Broad phonemes classes and their features (V-voiced, UV-unvoiced)

Broad Phoneme Classes	Features					
	Voicing Information	ZCR	STE	Duration	Presence of burst	Strength of formants
Vowels	V	Low	High	Long	No	Strong F1
Nasals	V	Low	Medium	Medium	No	Weak F2
Plosives	V/UV	Medium	Low	Short	Yes	-
Fricatives	V/UV	High	Low	Long	Yes	-
Approximants	V	Low	High	Short	No	High F3
Silence	UV	Low	Low	-	No	-

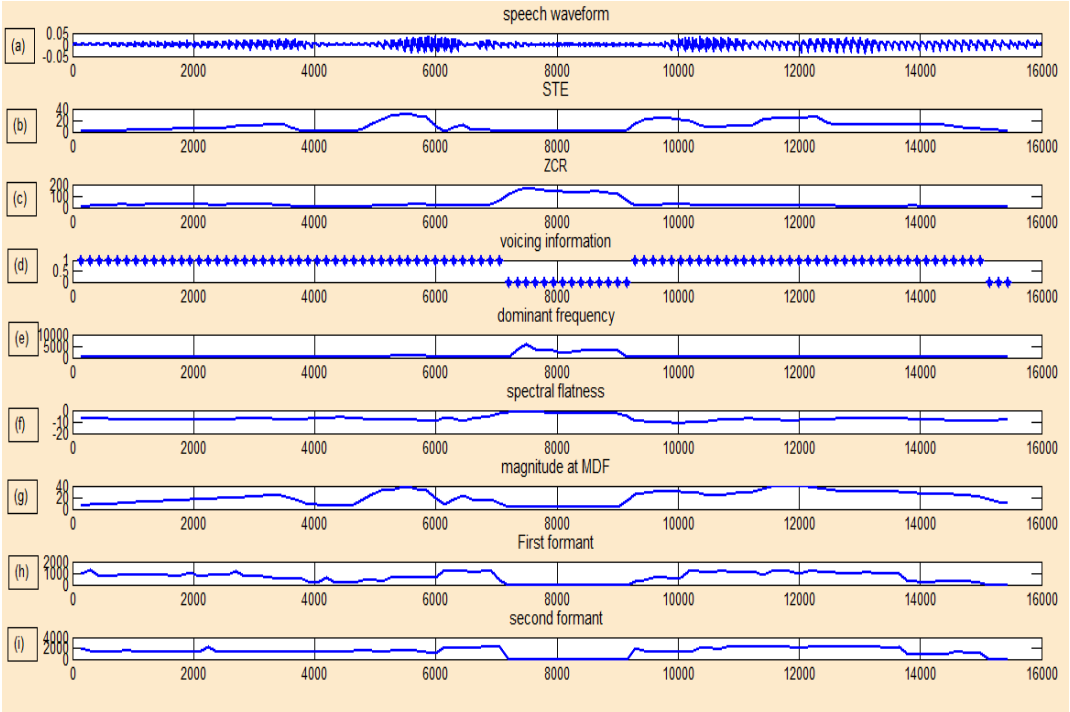

Figure 5. (a) Waveform, (b) STE, (c) ZCR, (d) voicing information, (e) MDF, (f) SF, (g) magnitude at MDF, (h) F1, and (i) F2 for the word //aakarshanam// in Malayalam.

2.3. Classifier Design

As explained, the feature vectors for each short frame (width 20 ms) are calculated and normalized. Then this features are applied to a classifier for broad phoneme classification. Here an multilayer feedforward artificial neural network [6][8][9] is used as a classifier. Neural networks are also similar to biological neural networks in performing functions collectively and in parallel by the units.

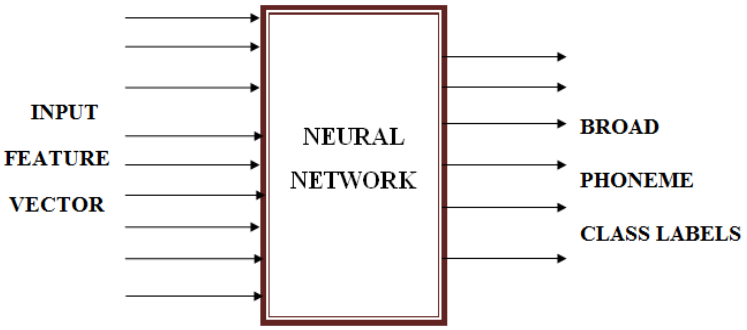

Figure 6. Neural network classifier for broad phoneme classification

In this work a feed-forward back propagation [9] network is created with five layers [8], one input layer, one output layer and three hidden layers. The final classifier has a structure 27L 54N 20N 10N 6L where L represents linear neurons and N represents non-linear neurons. Here the non-linear neurons use 'log sigmoid' activation function. Feed forward networks have one-way connections from input to output layers. Here the network training function used is 'traincgb', so that updates weight and bias values according to the conjugate gradient back propagation with Powell-Beale restarts.

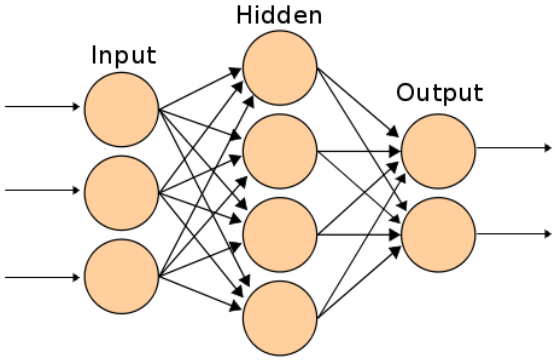

Figure 7. Feedforward neural network

3. EXPERIMENTAL RESULTS

3.1. Classifier Training

The training of the neural network is a time consuming step, which will end either by a validation stop or by reaching the maximum number of epochs. For training the network about 8 minutes of read speech is used. The speech data is divided into frames of size 20ms with 10ms overlap. For each of these frames 9 feature vectors were calculated, and along with this the difference between the previous frame and the succeeding frame are also considered. So all together 27 (9x3) inputs are applied to the network for each frame along with its manual transcription (supervised learning).

3.2. Evaluation

For testing, read speeches as well as some isolated words are collected. As done during training, the speech data is segmented into frames of 20ms with 10ms overlap. Feature vector of each frame is applied to the network classifier, so that the network outputs are interpreted as a label that suits the best for the given feature vector.

3.3. Results and discussion

The output labels obtained while testing the classifier using selected words are given below. One broad phoneme symbol is assigned for each short frame (20 ms frame size and 10 ms frame shift) of speech.

a. \\bhakshanam

MANUAL TRANSCRIPTION:

SSSSSSSSSSSPPPPPVVVVVVVVPPPPPPPPPFFFFFFFFFFFVVVVVVVVVNNNNNNVVVVV
VVVNNNNNNNNNNNNSSSSSSSSSSSSSS

OUTPUT OF PROPOSED METHOD:

SSSSSSSSSSSSPPPPVVVVVVVVPPPPPFFFFFFFFFFFFFPPVVVVVVVVVVVVVVVNVVVVV
VVVNNNNNNNNNNNNPPPPSSSSFSFSFS

b. \\haritham

MANUAL TRANSCRIPTION:

FFFFFFFFFFFFFFVVVVVVVVVVVAAAAVVVVVVVVPPPPPPPPPPVVVVVVNNNNNNN
NNNNNNSSSSSSSSSS

PREDICTED PHONEME LABELS USING THE PROPOSED METHOD:

SSFFFFFFFFFPVVVVVVVVVVVVVPPPAAAAAVVVVVVVPPPPPPVVVVVVVVVNNNVVV
NNNNPPNNNNNFFFFF

c. \\hasyam

MANUAL TRANSCRIPTION:

SSSSSSSSSSFFFFFFFFFFFVVVVVVVVVVVVVVVVVFFFFFFFFFFFFFFFFFFFAAAAAAVV
VVVNNNNNNNNNNNNNSSSSSSSSSSS

PREDICTED PHONEME LABELS USING THE PROPOSED METHOD:

SSSSSSSSSSSSPFFFFPPPVVVVVVVVVVVVVVVVVFFFFFFFFFFFFFFFFFFPPPAAAAAAVV
VVVVVVVVVVVVVNNNNNPPPSSSSSSS

d. \\manasantharam

MANUAL TRANSCRIPTION:

SSSSNNNNNVVVVVVVVVVVVVVNNNNNNNNNVVVVVVVVVFFFFFFFFFVVVVVVVVVVV
VNNNNNNNNPPPPVVVVAAAAVVVVVNNNNNNNNNNNNNNNNNN

PREDICTED PHONEME LABELS USING THE PROPOSED METHOD:

SNNNNNNNAAAAAVVVVVAAVVNNNNNNAAAAAVVVPFFFFFFFPPPAAAAAAAAVV
VVNNNNNVVVPAAAAVVVVNVAVVVVVNNNNNNNNNNNNNNNNNV

The confusion matrix obtained while testing some read speech is as shown in Table 3.

Table 3. Identification accuracy in percentage for proposed feature set, MFCC and combined system.

		IDENTIFICATION ACCURACY (IN %)																	
		USING PROPOSED FEATURES						USING MFCC						COMBINED SYSTEM					
		V	N	P	F	A	S	V	N	P	F	A	S	V	N	P	F	A	S
TEST 1	V	**53**	16	9	1	21	0	**51**	8	11	7	23	0	**57**	9	8	6	20	0
	N	20	**64**	12	0	4	0	9	**37**	25	5	23	1	12	**40**	22	4	22	0
	P	16	22	**47**	3	4	8	8	8	**39**	22	22	1	10	11	**38**	20	19	2
	F	0	0	16	**78**	1	5	1	0	13	**80**	6	0	0	0	5	**95**	0	0
	A	22	33	29	0	**16**	0	24	10	19	6	**41**	0	29	13	16	7	**35**	0
	S	4	2	14	4	0	**76**	0	2	27	20	2	**49**	0	2	22	9	2	**6**
TEST 2	V	**64**	17	4	0	14	1	**65**	11	9	3	12	0	**72**	7	5	3	12	1
	N	20	**57**	17	1	2	3	8	**65**	10	10	7	0	9	**67**	8	9	6	1
	P	16	34	**38**	2	3	7	11	6	**43**	25	15	0	14	6	**40**	25	14	1
	F	2	4	20	**74**	0	0	0	0	15	**83**	2	0	0	0	11	**87**	2	0
	A	37	26	25	2	**10**	0	27	7	15	15	**36**	0	32	8	15	15	**30**	0
	S	4	0	12	17	0	**67**	0	1	13	30	2	**54**	0	0	8	24	1	**6**
TEST 3	V	**61**	9	3	3	24	0	**30**	32	10	1	27	0	**51**	21	6	0	22	0
	N	43	**35**	0	0	22	0	11	**73**	13	0	3	0	19	**60**	16	0	5	0
	P	29	6	**32**	19	8	6	10	25	**41**	5	19	0	11	22	**40**	9	16	2
	F	22	0	13	**65**	0	0	0	13	39	**44**	4	0	0	13	26	**57**	4	0
	A	63	8	2	0	**27**	0	6	51	8	0	**35**	0	29	30	6	0	**35**	0
	S	0	0	25	0	0	**75**	0	3	56	6	9	**26**	0	3	47	6	9	**3**

Table 3 shows results obtained from three systems: system with proposed feature set, system using standard MFCC features and a score level combination/fusion system. Test 1 & 2 are male speech and TEST 3 is female speech. Here V, N, P, F, A, S represents Vowel, Nasal, Plosive, Fricative, Approximant, Silence respectively. Test 1, Test 2 and Test 3 are three different read speech data used for testing. Notice that the proposed system maps V, N, P, F, S almost correctly. But it misclassifies approximants as vowels. System using MFCC classifies approximants (A) more accurately. So we made a score level fusion of the proposed and MFCC classifier to get a combined system as shown in Figure 9.

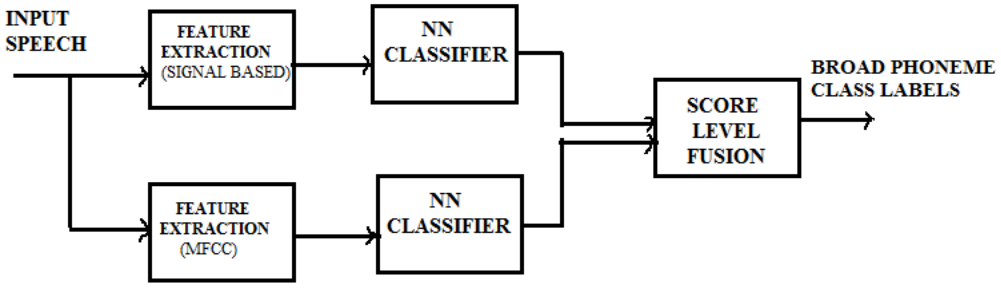

Figure 9. Combined system

With a combined system we have about 54%, 61%, and 46% of classifier accuracy for TEST 1, TEST 2 and TEST 3 respectively.

3.4 Application of proposed system for syllable structure prediction

For predicting the syllable structure from the frame level broad phoneme classifier output, some smoothing strategies are employed. For classes with long duration, like vowels, if there are more than 5 consecutive frames with label 'V', we accept the label 'V' to the syllable structure, else it is omitted. But for short phonemes such as plosives, occurrence of three or consecutive 'P' symbol is accepted as label to the syllable structure. Syllable structure prediction corresponding to three words as per this strategy is given in Table 4.

Table 4. Syllable structure prediction

Word	Actual syllable sequence	Predicted syllable sequence
//Hasyam//	/S/,/FV/,/FAVN/,/S/	/S/,/FV/,/FAVN/,/S/
// Haritham//	/FV/,/AV/,/PVN/,/S/	/FV/,/AV/,/PVN/,/F/
// Bhakshanam//	/S/,/PV/,/PFV/,/NVN/,/S/	/S/,/PV/,/PFV/,/VNP/,/S/

4. SUMMARY AND SCOPE OF FUTURE WORK

In this work, broad phoneme classification is attempted using signal level features. For broad phoneme classification, the features such as voicing, nasality, frication, vowel formants etc. are studied. From this study, we have come up with a set of feature vectors capable of performing broad classification of phonemes. Classifier is developed using feed forward neural network in order to automatically label each frame in terms of broad phoneme classes. The effectiveness of the proposed feature set is evaluated on speech database and is compared with standard MFCC features. Complimentary nature of both proposed features and MFCC is illustrated by combining the scores of the systems to gain an improvement in classification accuracy. Output of this classifier is used for broad syllable structure prediction.

The study can be further extended to phoneme classification instead of broad phoneme classification with an extended feature set. There is a scope for finer syllable structure prediction which has applications in area of automatic speech recognition and language recognition.

ACKNOWLEDGEMENTS

The authors would like to thank Department of Electronics and Information Technology, Government of India for providing financial assistance and Prof. B. Yegnanarayana, Professor, International Institute of Information Technology, Hyderabad for the motivation to carry out the work discussed in the paper.

REFERENCES

[1] G Douglas O' Shaughnessy, (2000) Speech communications-Human and Machine, IEEE press, Newyork, 2nd Edition.
[2] L Rabiner & B H Juang, (1993) Fundamentals of Speech recognition, Prentice Hall.
[3] Sadaoki Furui, “50 Years of Progress in Speech and Speaker Recognition Research”, ECTI Transactions on computer and Information technology, ol.1, No. 2, Nov 2005, pp. 64-74.
[4] Carol Y Espy (1986) “A Phonetically Based Semivowel Recognition System”, ICASSP 86, Tokyo.

[5] Carol Y. Epsy-Wilson, "A feature-based semivowel recognition system", J. Acoustical Society of America, vol. 96, No. 1, July 1994.pp.65-72.
[6] Jyh-Shing Roger Jang, Chuen- Tsai Sun & Eiji Mizutani, (1997) Neuro-Fuzzy and Soft Computing, Prentice Hall, 1st Edition.
[7] T. Pruthi, C. Y. Epsy-Wilson, "Acoustic parameters for automatic detection of nasal manner", Elsevier, Speech Communication 43 (2004), pp. 225-239.
[8] Fu Guojiang, "A Novel Isolated Speech Recognition Method based on Neural Network",2nd International Conference on Networking and Information technology, IPCSIT, vol. 17 (2011), IACSIT Press, Singapore, pp. 264-269.
[9] Caltenco F, Gevaert W, Tseno G, Mladenov, "Neural Networks used for Speech Recognition", Journal of Automatic Control, University of Belgrade, vol. 20, 2010.
[10] Sakshat Virtual Labs, Department of Electronics and Electrical Engineering, IIIT Guwahati.
[11] Speech production mechanism (Tutorial): Speech Signal Processing: Computer Science & Engineering: III Hyderabad Virtual Lab.
[12] Wavesurfer User Manual, http://www.speech.kth.se/wavesurfer/man18. html, pp:1-6, 9/23/2013.

Authors

Deekshitha G. graduated from Cochin University of Science and Technology in Electronics and Communication Engineering in 2012. She is currently doing masters degree in Advanced Communication and Information Systems at Rajiv Gandhi Institute of Technology, Kottayam Kerala, India. Her areas of interest are image processing and speech processing.

Leena Mary received her Bachelor's degree from Mangalore University in 1988. She obtained her MTech from Kerala University and Ph.D. from Indian Institute of Technology, Madras, India. She has 23 years of teaching experience. Currently she is working as Professor in Electronics and Communication Engineering at Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India. Her research interests are speech processing, speaker forensics, signal processing and neural networks. She has published several research papers which includes a book on Extraction and Representation of Prosody for Speaker, Speech and Language Recognition by Springer. She is a member of IEEE and a life member of Indian Society for Technical Education.