ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0162

RECOGNITION OF TAMIL SYLLABLES USING VOWEL ONSET POINTS WITH PRODUCTION, PERCEPTION BASED FEATURES

S. Karpagavalli[1] and E. Chandra[2]
[1]*Department of Computer Science, PSGR Krishnammal College for Women, India*
E-mail: karpagavalli@psgrkc.com
[2]*Department of Computer Science, Bharathiar University, India*
E-mail: crcspeech@gmail.com

Abstract

Tamil Language is one of the ancient Dravidian languages spoken in south India. Most of the Indian languages are syllabic in nature and syllables are in the form of Consonant-Vowel (CV) units. In Tamil language, CV pattern occurs in the beginning, middle and end of a word. In this work, CV Units formed with Stop Consonant – Short Vowel (SCSV) were considered for classification task. The work carried out in three stages, Vowel Onset Point (VOP) detection, CV segmentation and classification. VOP is an event at which the consonant part ends and vowel part begins. VOPs are identified using linear prediction residuals which provide significant characteristics of the excitation source. To segment the CV units, fixed length spectral frames before and after VOPs are considered. In this work, production based features, Linear Predictive Cepstral Coefficients (LPCC) and perception based features, Perceptual Linear Predictive Cepstral Coefficients (PLP) and Mel Frequency Cepstral Coefficients (MFCC) are extracted which are used to build the SCSV classifier using multilayer perceptron and support vector machine. A speech corpus of 200 Tamil words uttered by 15 native speakers was used, which covers all SCSV units formed with Tamil stop consonants (/k/, /ch/, /d/, /t/, /p/) and short vowels (/a/, /i/, /u/, /e/, /o/). The classifiers are trained and tested for its performance using predictive accuracy measure. The results indicate that perception based features, MFCC and PLP provides better results than production based features, LPCC and the model built using support vector machine outperforms.

Keywords:

Syllables, Consonant-Vowel Unit, Vowel Onset Point, Multilayer Perceptron, Support Vector Machine

1. INTRODUCTION

In speech recognition systems, to recognize the given speech utterance, it can be segmented into sub-word units and labeled with the help of effective sub-word unit recognizers. There are various sub-word units like phoneme, syllable and tri-phone. Sub-word model based approach is highly efficient in the development of vocabulary independent speech recognition systems. In Indian languages, the Consonant-Vowel (CV) segments occur with high frequency and those are the basic units of speech production. CV units can be considered as sub-word units [1], [2].

Tamil is an ancient Indian language spoken widely in the southern state of India, Tamil Nadu. Tamil is a syllabic language. There are 18 consonants and 12 vowels of different categories are present in Tamil language. Also one unique letter ‘aytham’ pronounced as guttural ‘k’, ‘g’ or ‘h’ appended to the preceding vowel, which occurs only between a short vowel and one of the hard class consonant [3]. The syllables are derived from a combination of 18 consonants and 12 vowels, a total of 216 different characters with unique symbols. Consonants are usually written with a dot on the top of the symbol as in க் (/k/), ச் (/ch/), when it combined with vowels to form syllables, dot on it is removed and a secondary symbol representing the vowel is included as க் (/k/) + அ (/a/) → க (/ka/) and க் (/k/) + இ (/e/) → கி (/ki/). The 18 consonants in Tamil language can be further classified as stops, nasals, laterals, trills and glides. In the proposed work, a subset of CV units formed with stop consonants + short vowels are considered, i.e., the work carried out as a 25 SCSV unit classification task since there are only 5 SCs and 5 SVs in Tamil language.

There are 5 stop consonants in Tamil language. Those pure stop consonants with their place of articulation are presented in Table.1.

Table.1. Stop Consonants in Tamil Language

Stop Consonant	Place of articulation
க் (/k/)	Velar
ச் (/ch/)	Palatal
ட் (/d/)	alveolar retroflex
த் (/t/)	Dental
ப் (/p/)	Bilabial

Table.2. Short Vowels in Tamil language

Short Vowels	Place of Articulation
அ (/a/)	Mid
இ (/i/)	Front
உ (/u/)	Back
எ (/e/)	Front
ஒ (/o/)	Back

There are 5 short vowels in Tamil language. Vowel sounds are produced with the tongue placed in different parts of the mouth. When pronouncing the vowels இ, எ, அ, ஒ, உ (/i/, /e/, /a/, /o/, /u/) tongue moves from back part of the mouth to the front. Based on the tongue positions, the vowels இ (/i/) and எ (/e/) are named front vowels, அ (/a/), a mid-vowel and ஒ (/o/) and உ (/u/) back vowels [4]. Those short vowels with their place of articulation information are tabulated in Table.2. Consonants are added with vowels to form the syllables. List of syllables formed with stop consonants and short vowels in Tamil language are listed in Table.3.

Table.3. Stop Consonant – Short Vowel (SCSV) units in Tamil language

Stop Consonants (SC)	Short Vowels (SV)				
	அ	இ	உ	எ	ஒ
	SCSV Units				
க்	க	கி	கு	கெ	கொ
ச்	ச	சி	சு	செ	சொ
ட்	ட	டி	டு	டெ	டொ
த்	த	தி	து	தெ	தொ
ப்	ப	பி	பு	பெ	பொ

This paper is organized as follows: In section 2 the review of literature which covers similar work carried out for different Indian languages by adopting various features and methods. Section 3 describes the proposed framework with clear sketch of the various stages in the work. Section 4 describes the detection of vowel onset point using Hilbert envelope of the LP residual convolved with a modulated Gaussian window. Section 5 presents the SCSV unit segmentation from the given speech utterance. Section 6 describes the LPCC, PLP, MFCC feature extraction process. Section 7 discusses the various classification algorithms adopted for SCSV unit classification. Section 8 presents the experimental results and comparative analysis of the classifiers performance. Section 9 provides the conclusion and the scope for future work.

2. RELATED WORK ON INDIAN LANGUAGES

Some of the CV unit recognition work carried out in Indian languages are studied in detail and elaborated as follows. Neural network models were developed for spotting Stop-Consonant-Vowel (SCV) segments in continuous speech [5]. Gangashetty et al. proposed CV unit recognition for isolated Hindi utterances using Auto Associative Neural Networks (AANN) [6]. CV classes are sub-grouped and different classifiers are modeled based on manner of articulation, place of articulation and vowels. The work is extended [7] for spotting multilingual CV units in speech using neural network models. They used frequently occurring CV units of three Indian languages Tamil, Telugu and Hindi.

Vuppala et al. [8] experimented CV unit recognition task on Telugu speech data under background noise using temporal and spectral processing methods. Recognition of vowel category of CV unit and then consonant category of CV unit is done separately and SVM, HMM hybrid models are employed to improve the performance. Thasleema et al. [9] carried out Malayalam CV unit classification using different classifiers like support vector machine with Decision Directed Acyclic Graph learning architecture (DDAGSVM), K-Nearest Neighbour and artificial neural network. They classified 5 classes of CV units which are unaspirated, aspirated, approximants, nasals and fricatives using hybrid wavelet based features which work better on noisy speech data. Anantha Natarajan et al. proposed an approach in which continuous speech is segmented into smaller speech units and each unit is classified either consonant or vowel using the formant frequencies on Tamil broadcast data [10].

3. PROPOSED WORK

The main objective of the work is to investigate that the consonant–vowel units can be used to build Tamil speech recognition systems rather using the mostly used sub-word unit phoneme. The work was carried out by considering only the SCSV units given in Table.3, and it involves various phases as depicted in Fig.1. In the VOP detection phase, the positions of VOP frame is identified for the given isolated Tamil word using the excitation information. In next phase, region around VOP frame of 90 to 110msec duration which contains the necessary acoustic characteristics of the SCSV unit is segmented. Then for each segment, LPCC, PLP, MFCC features are extracted and for each type of features extracted separate classifiers are built. The models are tested for its effectiveness using unknown test dataset.

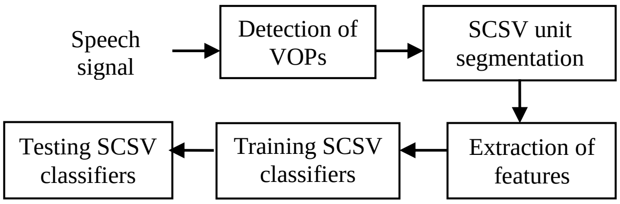

Fig.1. Block Diagram of the Proposed Framework

4. DETECTION OF VOPs

Vowel onset point is a point at which the consonant region ends and vowel region begins in a CV utterance. Utterances of CV units consist of different speech production events like closure, burst, aspiration, transition and vowel [11], [12]. All CV units have a distinct VOP in their production which is the significant property useful in CV unit segmentation or classification.

There are various methods to detect VOPs. Multilayer feed forward neural network (MLFFNN) model was proposed [6] to detect VOP by using the trends in the speech signal parameters at VOPs. In their approach two frames at pre-VOP region, VOP region and post-VOP region are considered and signal energy, residual energy and spectral flatness parameters and its ratio are used to form the input vector. The constructed network had three output nodes, pre-VOP, VOP, post-VOP. The trained model is used to identify the VOPs for which point of input, the VOP output node is maximum.

In another method to detect VOPs, auto associative neural network (AANN) model had been proposed [147]. They used five layered AANN with compression layer in middle and explored the distribution capturing of feature vectors. Similar work with little variation has been tried using AANN model to hypothesize the consonant and vowel regions and to detect VOPs in continuous speech. Detection of VOPs from noisy speech was proposed by Vuppala et al. [13] which exploit the spectral energy at formant frequencies of the speech segment present in glottal closure region. The formants are extracted by using group delay function and glottal closure instances are extracted using zero frequency filter based method.

In this work, detection of VOPs in CV utterance has been done using excitation information [14]. All CV utterances are processed by linear prediction analysis to extract the LP residual which carries the excitation information. In LP analysis, the dependencies among adjacent samples are estimated and then removed from the speech signal to obtain the residual signal. The prediction of current sample as a linear combination of past p samples from the basis of linear predictive analysis when p is the order of prediction. The predicted sample $\hat{s}(n)$ can be written as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k . s(n-k) \quad (1)$$

where, a_k are the linear prediction coefficients.

The difference between the envelope sample and the predicted sample is the prediction error $e(n)$ can be written as,

$$e(n) = s(n) - \hat{s}(n) \quad (2)$$

The Eq.(2) can be written as,

$$e(n) = s(n) + \sum_{k=1}^{p} a_k . s(n-k) \quad (3)$$

In the frequency domain, the Eq.(3) can be written as

$$E(Z) = S(Z) + \sum_{k=1}^{p} a_k . S(Z) Z^{-k} \quad (4)$$

i.e.

$$A(Z) = \frac{E(Z)}{S(Z)} = 1 + \sum_{k=1}^{p} a_k . Z^{-k} \quad (5)$$

Hilbert transform of LP residual $e_h(n)$ is calculated using the residual $e(n)$. The analytical signal $x(n)$ is given by

$$x(n) = e(n) + j e_h(n) \quad (6)$$

where, $e_h(h)$ is the Hilbert transform. The magnitude of the complex analytic signal of Eq.6 is called the Hilbert envelope of the signal $h(n)$ which can be written as,

$$h(n) = \sqrt{e^2(n) + e_h^2(n)} \quad (7)$$

Significant change in the amplitude of the Hilbert envelope of the LP residual is the clue to identify VOPs. A modulated Gabor window function [15] is convolved with $h(n)$ to find the VOP evidence plot.

$$g(n) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{n^2 \cos(\omega n)}{2\sigma^2}} \quad (8)$$

VOP evidence plot for a sample Tamil word ‘petti’ with two VOPs is shown in Fig.2.

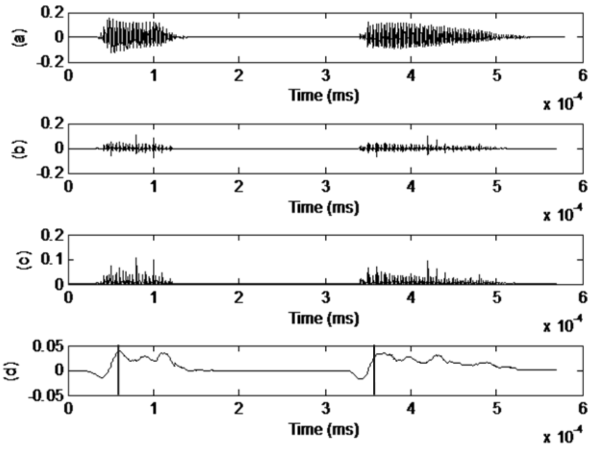

Fig.2. (a) Speech signal of sample Tamil word ‘petti’ with two SCSV units /pe/, and /ti/ (b) LP residual (c) Hilbert envelope of LP residual (d) VOP evidence plot with hypothesized VOP

From the VOP evidence plot, locations of all peaks preceded with negative region known as valleys are identified. A peak picking method is adopted to select peaks only with deep valleys and eliminate other spurious peaks which are closer to other peaks with distance less than 50 to 65 milliseconds. After choosing the peaks, the frames in which the VOPs occur are identified and used to perform SCSV unit segmentation. The complete procedure adopted for VOP frames identification is given below:

Step 1: Read the speech utterance

Step 2: Perform padding of zeros in both ends of the signal according to frame size

Step 3: Pre-emphasize the signal with $\alpha = 0.97$

Step 4: Compute LP residual using 20ms frame size with 10ms overlapping, LP order as 10

Step 5: Compute Hilbert Envelope of the LP residual

Step 6: Find the VOP evidence by convolving the Hilbert Envelope with modified Gaussian window of size 800

Step 7: Identify all peaks preceded with negative regions known as valleys

Step 8: Choose peaks only with deep valleys, else eliminate the peak

Step 9: Eliminate the spurious peaks which are closer to peaks with distance less than 50ms to 65ms

Step 10: Find the frame in which the selected peaks occur and label them as VOP frames

5. SCSV UNIT SEGMENTATION

In the proposed work, SCSV units are classified using discriminative models multilayer perceptron and support vector machine. As the models can process fixed dimensional patterns, it is important to devise a method to represent distinct SCSV utterances by fixed dimensional pattern. To classify each SCSV class, the region before VOP as corresponding to the manner of articulation, the transition region after VOP to the place of articulation and the steady vowel region provide useful information [6], [16]. The acoustic characteristics of each region influences the other regions, hence all regions are combined together to form a fixed length vector [17].

The acoustic characteristics of the SCSV unit are captured by performing short-time analysis of SCSV speech signals. SCSV utterances are represented as sequence of frames of size 20msec with a shift of 10 msec. For each frame 13 LPCC features, 12 PLP features and 13 MFCC features are extracted and each extraction method is described in section 6. The average duration of SCSV segment is 100 msec. As the beginning and end part of segments are affected by co-articulation effects, those are skipped. As frames around VOP carries significant information to distinguish SCSV units, fixed number of frames are considered. In this work, experiment is carried out for three different fixed length overlapping frames, 8 frames (4 frames before VOP frame + VOP frame + 3 frames after VOP frame), 9 frames (4 frames before VOP frame + VOP frame + 4 frames after VOP frame), and 10 frames (5 frames before VOP frame + VOP frame + 4 frames after VOP frame) for each SCSV unit.

6. FEATURE EXTRACTION

Speech based feature extraction techniques can be categorized as temporal analysis and spectral analysis methods. In temporal analysis the speech waveform itself is used wherein spectral analysis spectral representation of speech signal is used for analysis. Another challenge in speech signal is its variability which can be reduced by going for short-time spectrum analysis [18]. Selection of feature extraction methods is very vital because it must capture the important information from speech signal and suitable for the task. In the proposed work, LPCC - a production based feature extraction method and PLP, MFCC - perception based feature extraction methods are used to extract features from speech utterances. The steps involved in each feature extraction method are elaborated below.

6.1 LINEAR PREDICTIVE CEPSTRAL COEFFICIENTS (LPCC)

Linear predictive cepstral coefficients are derived from linear predictive coefficients, is the mostly used features in speech recognition tasks and works better than filter bank coefficients. LPC provides a good model of the speech signal and a good approximation to the vocal tract spectral envelope. LPC analysis of speech signal leads to a reasonable source-vocal tract separation in turn; a parsimonious representation of the vocal tract characteristics becomes possible. LPC is mathematically precise and easy to implement [19].

The speech signal is first pre-emphasised using a first order finite impulse response filter with pre-emphasis coefficient α = 0.97. The pre-emphasised speech signal is subjected to framing and windowing operation with frame duration of 20ms, frame shift of 10ms using hamming window. In auto correlation analysis, each frame of the windowed signal is auto correlated and provides $p + 1$ auto correlations for each frame, where p is the order of LPC analysis, typical value for p are 8 to 16. In LPC analysis step, each frame of $p + 1$ auto correlations are converted into LPC parameter set in which might be the LPC coefficients, the reflection coefficients, the log area ratio coefficient, the cepstral coefficients or any desired transformation of the above sets. The formal method for converting from autocorrelation coefficient to an LPC parameter set is known as Durbin's method. Then the LPC parameter set is converted into LP cepstral coefficients and the process is clearly depicted in Fig.3.

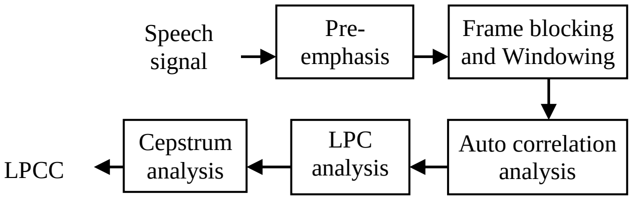

Fig.3. LPCC Feature Extraction

6.2 PERCEPTUAL LINEAR PREDICTIVE COEFFICIENTS (PLP)

A perceptual linear predictive coefficient is more robust than LPCC, based on short-term spectrum and includes perceptual aspects. Unlike LP analysis, it modifies the short-term spectrum of the speech by several psychophysically based transformations. PLP provides low dimensional representation of speech and highly related to the human processing of the signal. The power spectrum of the speech signal is converted to Bark scale which is similar to the human ear's perceptual model [20], [21]. In PLP coefficients calculation, initially, power spectrum estimate for the analysis window is computed i.e., the speech signal is passed through a hamming window and the squared magnitude of the FFT is calculated. Then the power spectrum is integrated within the critical band filter responses. In PLP, trapezoidal shaped filters are applied at roughly one-Bark intervals where the Bark axis is derived from the Eq.(9).

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right) \tag{9}$$

where, Ω represents the angular frequency in Bark scale, and ω represents angular linear frequency = $2\pi f$. This reduces the frequency sensitivity over the original spectrum estimate at high frequencies in particular as the bandwidth is high at high frequencies. Next equal loudness pre-emphasis is done to compensate non-equal perception of loudness at different frequencies using the Eq.(10),

$$E(\omega) = \frac{\left(\omega^2 + 56.8*10^6\right)\omega^4}{\left(\omega^2 + 6.3*10^6\right)\left(\omega^2 + 0.38*10^9\right)\left(\omega^6 + 9.58*10^{26}\right)} \tag{10}$$

where, ω represents angular linear frequency = $2\pi f$ and $E(\omega)$ is the energy at frequency ω. Then cube root is taken in the place of log as the perceived loudness is approximately the cube root of the intensity. As the log is not computed, the results are more like autocorrelation coefficients. IDFT of the power spectrum will result in autocorrelation coefficients. In the final step, autoregressive model is used to smooth the compressed spectrum. The autoregressive coefficients can be converted to cepstral variables. The block diagram for PLP feature extraction is shown in Fig.4.

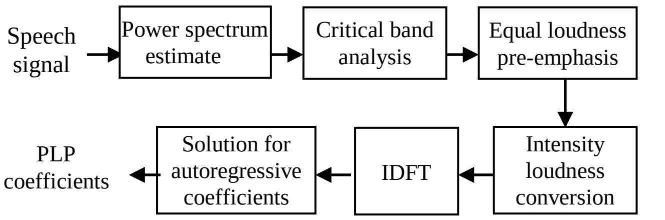
Speech signal
Power spectrum estimate
Critical band analysis
Equal loudness pre-emphasis
Intensity loudness conversion
IDFT
Solution for autoregressive coefficients
PLP coefficients

Fig.4. PLP Feature Extraction

6.3 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

Mel frequency cepstral coefficients are the parameterisation of choice for many speech recognition applications. MFCCs are computed from digitized speech signal sampled at 16000Hz. The speech signal is first pre-emphasised using a first order finite impulse response filter with pre-emphasis coefficient α = 0.97. The pre-emphasised speech signal is subjected to framing and windowing operation with frame duration of 20ms, frame shift of 10ms using hamming window.

The short-time Fourier transform analysis performed after windowing to compute magnitude spectrum. It is followed by filter bank design with triangular filters uniformly spaced on the mel scale between 300Hz to 3400Hz as lower and upper frequency limits. The filter bank is applied to the magnitude spectrum values to produce filter bank energies 20 per frame. Log-compressed FBEs are then de-correlated using the discrete cosine transform to produce cepstral coefficients.

The co-efficients are rescaled to have similar magnitude achieved through liftering with L = 22. The steps for MFCC feature extraction are shown in Fig.5.

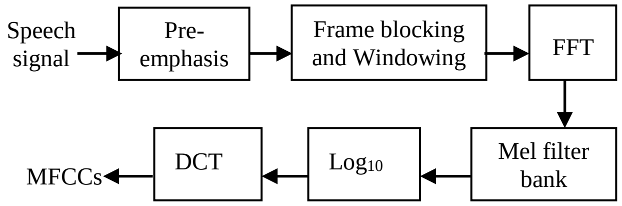

Fig.5. MFCC feature extraction

7. CLASSIFICATION ALGORITHMS

There are many machine learning algorithms used to solve the classification problems. This work employs multilayer perceptron and support vector machine to classify Tamil SCSV units.

7.1 MULTILAYER PERCEPTRON

The multilayer perceptron (MLP) is feed-forward artificial neural network, the most common neural network. The architecture of the MLP is as shown in Fig.6. It consists of an input layer, one or more hidden layers and an output layer. The number of hidden layers can be changed depending on problem data under training process. The output nodes can also be changed depending on classification of target output.

All inter-node connections have associated weights which are usually randomized at the initial step of training. The steps for computing the output of a single neuron are as follows: (1) compute the weighted sum of inputs to the neuron (2) add the bias to the sum (3) feed the sum as an input to the activation function of the neuron. The output of the activation function is defined to be the output of the neuron [22], [23].

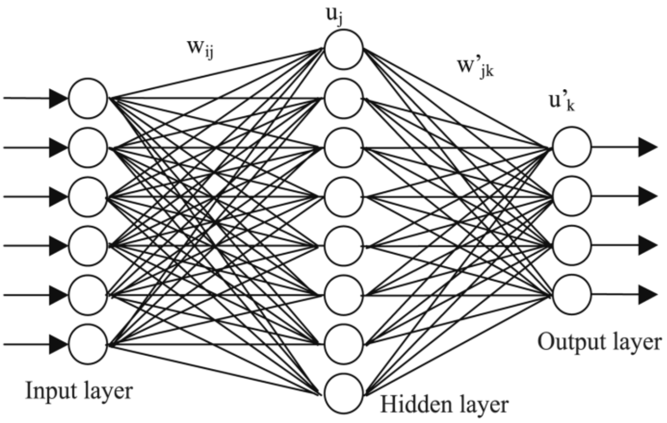

Fig.6. MLP architecture

The activation function with non-linear characteristics is important to discriminate the complex relationship in the feature space. A common activation function in the nodes of the hidden or output layer is the sigmoid function. However, other functions such as hyperbolic tangent function or quadratic function can also be employed. The activation function can be the same for all layers or a different function can be employed for each layer.

The training procedure, back propagation algorithm employs the method of gradient descent, which tends to minimize the mean squared error between the output of an MLP network and the desired output. The back-propagation algorithm compares the result that is obtained with the result that was expected and computes the error, calculates delta value in backward direction. Then computes delta weight to adjust the weights and repeats the training. In iterations, the error decreases and the neural model gets closer and closer to produce the desired output.

7.2 SUPPORT VECTOR MACHINES

Support vector machines are a popular learning method that can be used to build a complex classification model. The three key concepts behind SVM-based classification are margins, duality, and kernel functions. Assume that, a task requires the binary classification of m data points, each having classification labels y_i = ±1. Each data point is represented by a d-dimensional feature vector. The classification function to use that describes the discriminating plane is,

$$f(x) = sign(w.x - b) \tag{11}$$

The vector w describes the orientation of the plane and b describes the offset of the plane from the origin. Assuming the classes associated with the data points are linearly separable, there are an infinite amount of planes that will correctly classify the data.

To determine the maximum margin between the two classes is to maximize the margin between two parallel supporting planes. A plane supports a class if all points in the class are on one side of the plane. These supporting planes required that $w \cdot x_i - b \geq 1$ for class 1 and $w.x_i - b \leq -1$ for class -1. The distance between these planes is maximized to determine the optimal plane for classification, and the highlighted points are support vectors which are shown in Fig.7.

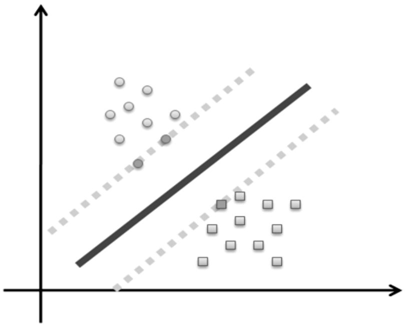

Fig.7. Support Vector Machine

The distance between the two supporting planes is $\frac{2}{\|w\|^2}$. Thus, maximizing the margin is equivalent to minimizing $\frac{\|w\|^2}{2}$, using the quadratic programming problem given in the following equations:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad (12)$$

$$s.t. w.x_i \geq b + 1 y_i \in \text{class } 1$$

$$w.x_i \leq b - 1 y_i \in \text{class } -1$$

The solution only depends on constructing the plane based upon these points. The equivalency of the two solutions illustrates the concept of duality and the Lagrangian dual.

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i \cdot x_j - \sum_{i=1}^{m}\alpha_i \quad (13)$$

$$s.t. \sum_{i=1}^{m} y_i \alpha_i = 0 \text{ and } \alpha_i \geq 0 i = 1,\ldots,m$$

Solving Eq.(11), Eq.(12) and Eq.(13) will yield the normal to the plane $w = \sum_{i=1}^{m} y_i \alpha_i x_i$. In practice, only a small subset of the α_i multipliers will be non-zero, and the corresponding x_i are the support vectors. SVM can be applied to linearly inseparable data where a slack variable is introduced to each constraint and then added as a weighted penalty term, as in Eq.(14),

$$\min_{w,b,z} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{1} z_i \quad (14)$$

$$s.t. y_i(w \cdot x_i - b) + z_i \geq 1 z_i \geq 0 i = 1,\ldots,m$$

where, C is the regularization parameter and Eq.(14) can be solved using the Lagrangian duality.

In the case of non-linear classification problems, kernel functions are used. Kernel function calculates a dot product between two vectors that have been (nonlinearly) mapped into a high dimensional feature space [24]-[26]. The final form of the quadratic programming problem is,

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{m}\alpha_i \quad (15)$$

$$s.t. \sum_{i=1}^{l} y_i \alpha_i = 0 \; C \geq \alpha_i \geq 0 \; i = 1,\ldots,m$$

Any kernel function can be substituted into this calculation without making any algorithmic changes. The popular kernel functions that are suitable for use in SVMs are linear, polynomial, radial basis function and sigmoid.

8. EXPERIMENT AND RESULTS

8.1 DATASET

A speech corpus of 15000 utterances was prepared which contains 200 Tamil isolated words uttered by 15 native speakers of Tamil and each speaker uttered each word 5 times. The words are selected on the basis of presence of more than one SCSV unit and to cover all SCSV units formed with Tamil stop consonants (/k/, /ch/, /d/, /t/, /p/) and short vowels (/a/, /i/, /u/, /e/, /o/). In this work, GoldWave digital audio editor and high quality microphone was used in recording the data. The recordings carried out in quiet room environment and at sampling rate 16kHz with 16 bit PCM.

8.2 EXPERIMENTAL SETUP

For each Tamil isolated word utterance, VOP frames are identified using excitation information. As frames around VOP carries significant information to distinguish SCSV units, fixed number of frames are segmented to represent SCSV units. In this work, experiment is carried out for three different fixed length overlapping frames, 8 frames (4 frames before VOP frame + VOP frame + 3 frames after VOP frame), 9 frames (4 frames before VOP frame + VOP frame + 4 frames after VOP frame), and 10 frames (5 frames before VOP frame + VOP frame + 4 frames after VOP frame) for each SCSV unit. LPCC, PLP, MFCC features of the fixed length pattern, i.e., 8 frames, 9 frames, 10 frames have been extracted used as pattern vector to train and test the classifiers.

As some of the SCSV units like /te/, /to/ etc., occurs minimally in the Tamil language, to build the SCSV classifier using MLP and SVM, a minimum of 100 instances of each SCSV unit is used. In multilayer perceptron, the network was designed with input layer, one hidden layer and 25 nodes in output layer. The network was trained with back-propagation algorithm with gradient descent learning function, sigmoid activation function. After training, the network was updated with new weights.

In support vector machine, models are built with linear, polynomial, RBF kernel with training data. Regularization parameter C is assigned different values in the range of 1 to 10 and found that the model performs better and reaches a stable state for the value of C = 8. The parameter settings used for polynomial, degree of polynomial d = 2 and for RBF kernel g = 0.2.

8.3 RESULT ANALYSIS

A total of 18 models were built using 2 classifiers MLP and SVM for a set of 3 different features and 3 different numbers of frames. The SCSV classifiers are tested with minimum 50 instances of each SCSV unit. Predictive accuracy can be defined as the ratio of number of correctly classified instances and total number of instances. It is used as a performance measure to analyze the performance of the built models with LPCC, PLP and MFCC features.

The performances of the models are summarized in Table.4, Table.5 and Table.6.

Table.4. Performance of the classifiers using LPCC

Classifier	**Recognition Accuracy in %**		
	8 Frames	**9 Frames**	**10 Frames**
MLP	75.4	76.2	78.0
SVM Linear kernel (C = 2)	65.2	70.4	71.6
SVM Linear kernel (C = 4)	71.8	71.8	72.0
SVM Linear kernel (C = 6)	72.4	73.0	73.2
SVM Linear kernel (C = 8)	73.0	74.3	74.5
SVM polynomial kernel	75.2	75.3	76.6
SVM RBF kernel	76.7	77.3	78.6

The results indicate that almost all the classifiers designed with 10 frames information recognize SCSV unit more accurately than 8 and 9 frames of acoustic information. The SVM classifier with RBF kernel performs substantially better compared to other models.

Table.5. Performance of the classifiers using PLP

Classifier	**Recognition Accuracy in %**		
	8 Frames	**9 Frames**	**10 Frames**
MLP	73.1	75.2	81.8
SVM Linear kernel (C = 2)	70.1	71.4	73.1
SVM Linear kernel (C = 4)	71.3	72.5	74.9
SVM Linear kernel (C = 6)	71.3	72.5	77.2
SVM Linear kernel (C = 8)	72.0	74.5	78.0
SVM polynomial kernel	78.2	81.3	83.9
SVM RBF kernel	81.7	82.5	85.7

The performance of the perception based feature PLP, shows that the SCSV models are providing better recognition accuracy compared to the production based feature LPCC. As well as the models built using 10 frames are able to classify the given SCSV unit better than other models using 8 and 9 frames of acoustic information.

Table.6. Performance of the classifiers using MFCC

Classifier	**Recognition Accuracy in %**		
	8 Frames	**9 Frames**	**10 Frames**
MLP	75.4	78.2	82.0
SVM Linear kernel (C = 2)	70.2	75.4	76.6
SVM Linear kernel (C = 4)	70.2	75.4	76.6
SVM Linear kernel (C = 6)	72.4	76.5	78.8
SVM Linear kernel (C = 8)	73.3	77.3	80.0
SVM polynomial kernel	80.2	82.3	84.6
SVM RBF kernel	91.7	92.5	94.4

The performance of the mostly used MFCC feature based classifiers provided in Table.6, indicates that the prediction accuracy rate is high compared to other models based on LPCC and PLP. Support vector machine-RBF kernel based models considerably better in prediction of SCSV units.

The overall results indicate that (i) to recognize each SCSV unit, 110 ms duration of speech signal information is sufficient, (ii) the perception based MFCC features carries most relevant spectral information to identify SCSV units (iii) SVM classifier with RBF kernel performs substantially better compared to other models used in the experiment.

9. CONCLUSION

Most of the state-of-the-art speech recognition systems are designed using phoneme as the sub-word unit. Since, Tamil language is syllabic in nature and Consonant-Vowel pairs form most of the syllables, this work was carried out with the focus that to confirm the higher level sub-word unit i.e. CV unit can be used to build Tamil speech recognition systems. In this work, an approach to classify stop consonant – short vowel (SCSV) units of Tamil language based on vowel onset points has been attempted. The models are built using multilayer perceptron and support vector machine using LPCC, PLP, MFCC features with different length pattern vector to recognize the given SCSV unit. In future, other approaches for VOP detection can be applied and the work can be extended to classify all CV units in Tamil language.

REFERENCES

[1] C.C. Sekhar, S.M. Santhosh and B. Yegnanarayana, “A Modular Approach for Recognition of Isolated Stop-Consonant-Vowel (SCV) Utterances in Indian Languages”, *Journal of the Acoustic Society of India*, Vol. 23, No. 1, pp. 28-35,1995.

[2] G. Lakshmi Sarada, A. Lakshmi, Hema A. Murthy and T. Nagarajan, “Automatic Transcription of Continuous Speech into Syllable-like Units for Indian Languages”, *Sadhana*, Vol. 34, No. 2, pp. 221-233, 2009.

[3] S. Ilakkuvanar, “*Tholkappiyam*”, (in English), Kural Neri Publishing House, 1963.

[4] K. Karunakaran and V. Jeya, "*moZhiyiyal*", Chennai: Kavitha Pathippakam, 1997.

[5] C.C. Sekhar and B. Yegnanarayana, "Recognition of Stop-Consonant-Vowel (SCV) Segments in Continuous Speech using Neural Network Models", *Journal of the Institution of Electronics and Telecommunication Engineers*, Vol. 42, No. 4-5, pp. 269-280, 1996.

[6] S.V. Gangashetty, C.C. Sekhar and B. Yegnanarayana, "Detection of Vowel Onset Points in Continuous Speech using Autoassociative Neural Network Models", *Proceedings of International Conference on Spoken Language Processing*, pp. 401-410, 2004.

[7] Suryakanth V. Gangashetty, C. Chandra Sekhar and B. Yegnanarayana, "Spotting Multilingual Consonant-Vowel units of Speech using Neural Network Models", *Nonlinear Analyses and Algorithms for Speech Processing*, Vol. 3817, pp. 303-317, 2005.

[8] Anil Kumar Vuppala1, K. Sreenivasa Rao and Saswat Chakrabarti, "Improved Consonant-Vowel Recognition for Low Bit-rate Coded Speech", *International Journal of Adaptive Control and Signal Processing*, Vol. 26, No. 4, pp. 333-349, 2011.

[9] T.M. Thasleema and N.K. Narayanan, "Wavelet Transform based Consonant-Vowel (CV) Classification using SVMs", *Proceedings of the 19th International Conference on Neural Information Processing*, pp. 250-257, 2012.

[10] V. Anantha Natarajan and S. Jothilakshmi, "Segmentation of Continuous Speech into Consonant and Vowel units using Formant Frequencies", *International Journal of Computer Applications*, Vol. 56, No. 15, pp. 24-27, 2012.

[11] S.R. Mahadeva Prasanna, Suryakanth V. Gangashetty and B. Yegnanarayana, "Significance of Vowel Onset Point for Speech Analysis", *Proceedings of International Conference on Signal Processing and Communications*, pp. 81-88, 2001.

[12] Dik J. Hermes, "Vowel Onset Detection", *The Journal of the Acoustical Society of America*, Vol. 87, No. 2, pp. 866-873, 1990.

[13] Anil Kumar Vuppala1 and K. Sreenivasa Rao, "Vowel Onset Point Detection for Noisy Speech using Spectral Energy at Formant Frequencies", *International Journal of Speech Technology*, Vol. 16, No. 2, pp. 229-235, 2013.

[14] S.R. Mahadeva Prasanna and B. Yegnanarayana, "Detection of Vowel Onset Point Events using Excitation Source Information", *Proceedings of 9th European Conference on Speech Communication and Technology*, pp. 1133-1136, 2005.

[15] D. Gabor, "Theory of Communication. Part 1: The Analysis of Information", *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, Vol. 93, No. 26, pp. 429-457, 1946.

[16] Peri Bhaskararao, "Salient Phonetic Features of Indian Languages in Speech Technology", *Sadhana*, Vol. 36, No. 5, pp. 587-599, 2011.

[17] K. Vuppala, K. Sreenivasa Rao and Saswat Chakrabarti, "Spotting and Recognition of Consonant-Vowel units from Continuous Speech using Accurate Vowel Onset Points", *Circuits, Systems, and Signal Processing*, Vol. 31, No. 4, pp. 1459-1474, 2012.

[18] M.A. Anusuya and S.K. Katti, "Front End Analysis of Speech Recognition: A Review", *International Journal on Speech Technology*, Vol. 14, No. 2, pp. 99-145, 2011.

[19] Lawrence Rabiner and Biing-Hwang Juang, "*Fundamentals of Speech Recognition*", New Jersey: Prentice Hall, 1993.

[20] Hynek Hermansky, "Perceptual Linear Predictive PLP Analysis of Speech", *Journal of the Acoustic Society of America*, Vol. 87, No. 4, pp. 1738-1752, 1990.

[21] Homayoon Beigi, "*Fundamentals of Speaker Recognition*", New York: Springer, 2011.

[22] Simon O. Haykin, "*Neural Networks and Learning Machines*", 3rd Edition, New York: Prentice Hall. 2009.

[23] S.N. Sivanandam, S. Sumathi and S.N. Deepa, "*Introduction to Neural Network using MATLAB 6.0*", Tata McGraw Hill, 1998.

[24] Koby Crammer and Yoram Singer, "On the Algorithmic Implementation of Multi-class Kernel-based Vector Machines", *The Journal of Machine Learning Research*, Vol. 2, pp. 265-292, 2001.

[25] H.W. Ian, "*Data Mining-Practical Machine Learning Tools and Techniques*", 2nd Edition, Elsevier, 2008.

[26] John Shawe-Taylor and Nello Cristianini, "*Support Vector Machines and other Kernel-based Learning Methods*", London: Cambridge University Press, 2000.