Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2022.022681
Article

Tech Science Press

Speech Quality Enhancement Using Phoneme with Cepstrum Variation Features

K. C. Rajeswari[1,*], R. S. Mohana[2], S. Manikandan[3] and S. Beski Prabaharan[4]

[1]Department of Computer Science and Engineering, Sona College of Technology, Salem, 636005, Tamil Nadu, India
[2]Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, 638060, Tamil Nadu, India
[3]Department of Information Technology, E.G.S. Pillay Engineering College, Nagapattinam, 611002, Tamil Nadu, India
[4]Department of Computer Science and Engineering, Chitkara University Institute of Engineering and Technology, Rajpura, 140401, Punjab, India
*Corresponding Author: K. C. Rajeswari. Email: rajeswarikc@sonatech.ac.in

Received: 15 August 2021; Accepted: 26 October 2021

Abstract: In recent years, Text-to-Speech (TTS) synthesis is taking a new dimension. People prefer voice embedded toys, online buyers are interested in interactive chat application in the form of text-to-speech facility, screen readers for visually challenged people, and many more applications use TTS module. TTSis a system that is capable of converting the arbitrary text input into natural sounding speech. It's success lies in producing more human like speech sounding more natural. The most importanttechnical aspect of TTS is feature extraction process. Both text and speech features are needed but it is not that easy to select meaningful and useful features from the text or from speech. There are many feature extraction techniques available for both text and speech, still there is a need for very simplest form of feature extraction technique. Though the emergence of Deep learning technique automates feature extraction, it is suitable only when the volume of data is enormous. This paper proposes a novel text and speech feature extraction technique which is based on special symbols present in the text and phoneme with cepstrum variation of the speech signal respectively. These techniques are simple and works well for real-time applications in which size of data is small or moderate. The proposed methods not only extract useful features but also meaningful features in terms of fetching the salient traits of the text and speech cepstrum. The experimental results have shown that the quality of speech is increased by 14% when compared to the other conevntional feature extraction techniques.

Keywords: Speech synthesis; tamil TTS; feature extraction; prosody; intonation

1 Introduction

Every language has its own grammar for the spoken and written form of text. Some languages may or may not differ in written and spoken form, but the spoken of Tamil differs from its written form. The written form of Tamil text is referred to as 'Senthamizh' and spoken form is referred as 'Iyatramizh'. Most of the scholarly books, media, public speaking, official and Government writings use Senthamizh whereas Tamil speaking community uses Iyatramizh. The language also has diglossia that refers to the same language being spoken in more than one form with respect to region or community. The research work focuses only on Senthamizh because it is the script form of the language. It is necessary to concentrate on the syntax of the text or sentence specific to the language because the Parts of speech information are the building blocks of the language. Makawana et al., in a survey mention that parsing is one of the important tasks in natural language processing (NLP) [1]. Parsing provides information about the syntactical structure of the sentence. Sudhakar et al., adopted forward parsing technique for sentiment analysis for which intonation patterns is of high concern [2]. For example, English sentence keeps the verb in the middle of the sentence whereas Tamil sentence keeps the verb phrase at the end of the sentence.

The general structure of Tamil sentence is subject-object-verb. The object is preceded by the subject and the sentence is concluded by the verb. A valid Tamil sentence can be formed in any one of the following ways:

Sentence formed only with a verb. Example: நடந்துவிட்டது is a verb stating “It has happened” in English. Sentence formed with a subject and an object but not verb. Example: இது என் நாடு–“This is my nation” in English, verb is present in the sentence. Sentence formed with a subject, an object and a verb. Example: ராமுபாடம் படித்தான்—“Ramu read the book” in English.

Since it is possible to form sentences without subject and object in Tamil, it is said to be a null-subject language. Tamil is a classical language and has undergone many transformations in spoken and written form over a period of 2,200 years. The written form of Tamil text is given by the oldest grammar book Tolkkappiyam ((தொல்காப்பியம்).). The present form of Tamil sentence structure has been prescribed by Nannul ((நன்னூல்).). The language is grammar rich and also agglutinative in nature. The word has one lexical root with or without suffixes. In spite of all these, any text will consists of many special characters or symbols. It is a very important element since it also conveys useful prosodic information and sometimes even pragmatic information. For example,

Statement 1 : எனக்கு நேற்று, மாலை கிடைத்தது. (Yesterday I got the garland)

Statement 2 : எனக்கு, நேற்று மாலை கிடைத்தது. (Yesterday evening, I got)

Though the sentences share the same set of words, they convey different meaning. The first statement conveys the message that “Yesterday I got the garland” whereas the second statement conveys “Yesterday evening I got”. In both the sentences, மாலை is a noun, but conveys different meaning. In statement 1: மாலை means garland and in statement 2: மாலை means evening. The placement of special symbol “,” is responsible for the change in meaning of the sentences. It is evident from the simple example that, to incorporate appropriate prosody in the speech, such special symbols must be considered.

The purpose of speech feature extraction differs for various applications. In speech recognition application, the recognition performance greatly depends on the features extracted and it is also responsible for computing the features in sequence to be stored in vectors. In speech synthesis application, the features are used to analyze the short term spectrum of the signal, so that the modifications in prosody can be incorporated according to the features extracted. Pooja et al., have discussed many speech feature extraction techniques namely LPC, LPCC, MFCC, PLP, FFT, RASTA, and DWT [3]. Rajeswari et al., discussed the transformation based MFCC speech feature extractions [4]. Shreya et al., have reviewed various speech feature extraction methods and highlighted the merits and demerits as listed in the Tab. 1 [5].

All the research contributions for speech generation so far analyzed only any one of the features such as Linear prediction (LP), Mel frequency cepstral coefficients (MFCC), Linear predictive cepstral coefficients (LPCC), Fast Fourier transform (FFT), Discrete wavelet transform (DWT) and Guassian model (GM). Each of the mentioned techniques suffers from one amongst the problems like lack of time information; requirement of high sampling, problems related to shifts and sensitive to noise. The proposed work has modified the approach of analysis as FFT based enhanced MFCC spectral feature analysis, DWT based enhanced MFCC spectral feature analysis and FrFT based enhanced MFCC spectral feature analysis. The proposed feature choice takes into account this information to model the intonation component. The research work proposes "Function of intonation" as a feature extracted from text. The proposed approach provides a clear picture about the work of interest, i.e., pitch track, peak variations, depth and frequency variations, spectrum analysis through which improvement in quality of synthesized speech is ensured. Phoneme with Cepstrum Variation (PCV) is the novel speech feature extraction technique proposed. Redundant Kernel Learning (RKL) is a kind of machine learning approach used to perform further learning process.

Table 1: Advantages and disadvantages of speech feature extraction techniques

Technique	Advantages	Disadvantages
Linear predictive coding	• Describes the time-varying linear system that effectively represents the functioning of vocal tract	• Similar sounds cannot be identified
	• Computes with good speed	• Hardly analyzes the local features since LPC assumes the speech signals are stationary
	• Requires less bit rate for transmission and hence less bandwidth	• Cannot produce high quality in speech since certain speech stays back as scum
		• The less bit rate naturally reduces the quality of speech
MFCC	• Responsible to capture important attributes of speech	• Performance depends on the number of filters
	• Complexity is very less	• Certainly MFCC gives inaccurate result in the presence of noise
RASTA	• Robustness high	• Error prone while filtering
	• Able to capture frequencies of low modulations	• Combination with PLP gives better performance
	• Capable of analyzing both slow and fast variations of speech	
PLP	• Clarity in perceptually observable components	• Best suitable for speech recognition and speaker verification systems alone
	• Spectral characteristics matches human auditory system	• Limited capability in dealing with distortions
	• No pre-processing is required to remove perceptually unimportant tonal components	

2 Proposed Text Feature Extraction

The Tamil language is grammar rich, agglutinative in nature, that makes a real challenge in extracting the features from the text. The existing feature extraction methods by Akshay et al., had extracted horizontal and vertical curves or lines from the written form of text [6]. This is possible only when the text is presented in the image. Ramu Reddy et al., have proposed a two stage intonation model using neural networks to predict the F0 values of the syllables in a text [7]. Rizi Ahmed et al., in an investigation, highlight the possible differences in intonation patterns among Persian speaker's English Interlanguage [8]. Romportl et al., have developed F0 model that depends on the position of the prosodic word present in the prosodic structure of the sentence [9]. The slope method and discrete wavelet transform method are the two most widely used methods to extract features from handwritten character and optical character. The proposed research work extracts the intonation function based features from the special symbols present in the text and the sentence is further classified as shown in the Tab. 2.

Table 2: Special symbols and sentence classification

Special symbol	Classification of the sentence
! (exclamation mark)	Exclamatory
? (question)	Interrogative
,(comma)-(hyphen) ;(semicolon) " "(quotation)	Declarative
None of the above	Others

Fig. 1 shows the design of proposed intonation model. The Tamil input text is analyzed to classify the sentence into declarative, interrogative, exclamatory or others. This is achieved with the help of symbols present in the sentence. According to the table, if the symbols comma, semicolon, hyphenation or quotation is present, then the sentence is classified as declarative sentence. If the sentence is declarative, then the complexity of the sentence is analyzed. If the number of words present in the sentence does not exceed 10 and if there is no long vowel present in the last two words then it is considered as simple sentence otherwise complex sentence. A declination in the F0 (pitch) will be a suitable contour for the simple sentence and rise in F0 will be a suitable contour otherwise. If a question mark is present in the sentence, then it is classified as interrogative sentence. If an exclamation mark is present in the sentence then it is classified as exclamatory sentence. If the sentence is exclamatory, then the position or place of exclamation is analyzed and the word previous to the symbol must be assigned a rise in F0 contour. If there is no symbol present in the sentence, it is considered to be in others category. It is checked for the presence of long vowel in the words to raise F0, otherwise declination in F0 will be appropriate.

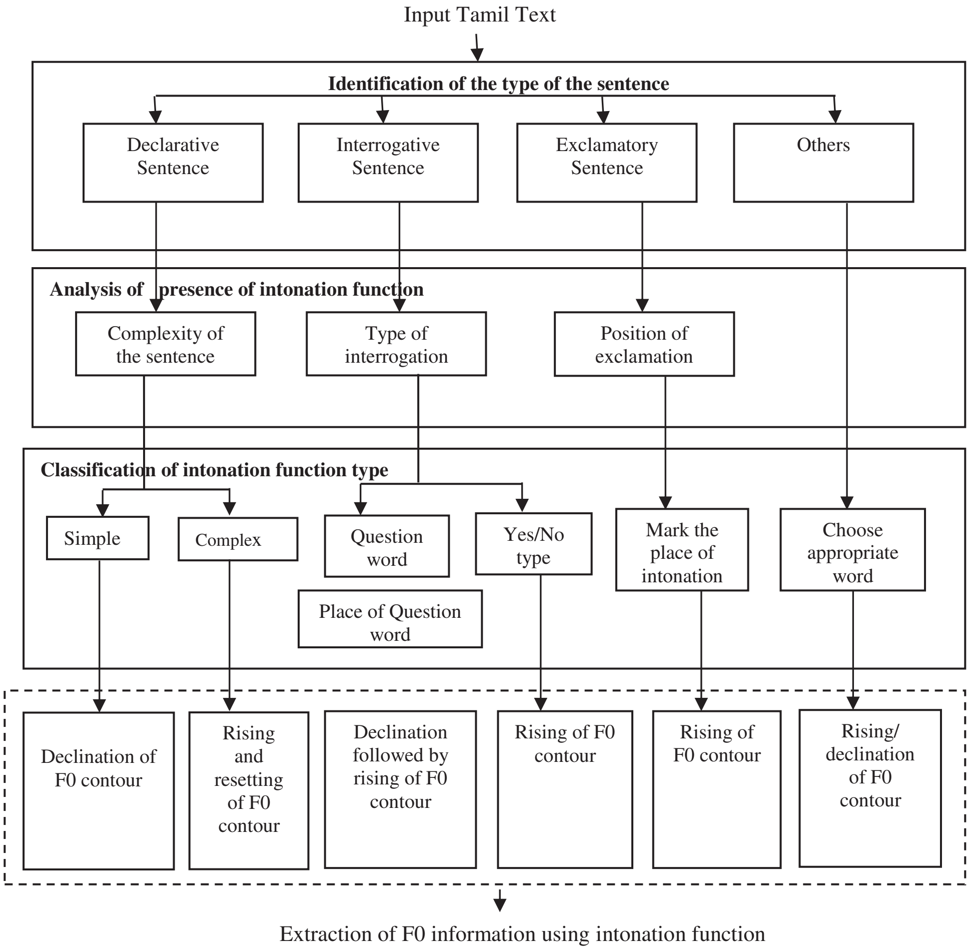

Input Tamil Text
Identification of the type of the sentence
Declarative Sentence
Interrogative Sentence
Exclamatory Sentence
Others
Analysis of presence of intonation function
Complexity of the sentence
Type of interrogation
Position of exclamation
Classification of intonation function type
Simple
Complex
Question word
Yes/No type
Place of Question word
Mark the place of intonation
Choose appropriate word
Declination of F0 contour
Rising and resetting of F0 contour
Declination followed by rising of F0 contour
Rising of F0 contour
Rising of F0 contour
Rising/ declination of F0 contour
Extraction of F0 information using intonation function

Figure 1: System design of text feature extraction

The intonation function includes attitude, grammar, focus, discourse, psychology and index terms along with the special symbols such as comma, semicolon, exclamation, question, hyphenation and quotation. The attitudinal intonation function is accountable for the emotions and attitude present in the phrase. The grammatical intonation function will clearly convey the difference in the structure of the phrase. The focus intonation function concentrates on the important elements present in the message. The discourse intonation function highlights the importance of subordinate clause rather than the main clause. The psychological intonation function gives importance to the perceivable units of the message. The indexical intonation function considers one's social identity in a group.

Pseudo code 1: Text Feature Extraction

```
INPUT:        Tamil text
OUTPUT:       Features
CHECK:        special symbol
FOR each sentence S
TOKENIZE Sentence to generate symbols
BEGIN
    FOR all symbols $s_i$ in a Sentence S
BEGIN
IS $s_i$ is a Special symbol then
DEFINE type $T_i$ ($s_i$ );
DEFINE Intonation type $In_i$ ($s_i$ );
ELSE
$s_i$ is a Long Vowel;
SEARCH (position of $s_i$);
DEFINE Intonation type $In_i$ ($s_i$);
END
END
```

The pseudocode 1 illustrates how the intonation function based features are extracted from the text. The Tamil text is parsed to calculate the length of the sentence and count the number of words present in the sentence. It is also checked for the presence of special symbols/characters. The position where special symbols are present is stored separately. The sentences are classified based on the special symbols that appear in the sentence. The labels such as 1, 2, 3 and 4 are assigned for each classification of the sentence type. After identifying the type of the sentence, it is further analyzed to identify the type of intonation function such as attitudinal, grammatical, focus, discourse, psychological and indexical to which the sentence belong.

3 Proposed Speech Feature Extraction

Researchers still strive to provide suitable speech feature extraction techniques. Shrawankar et al., has mentioned that spectrum obtained through FFT provides more useful information about the speech signal [10]. Shafi et al., has compared the features obtained by employing Discrete cosine transform and

Discrete Fourier Transform (DFT) in the process of removing the noise from the speech [11]. Kurzekar et al., has compared various feature extraction techniques for speech systems [12]. Vimal et al., has employed a Discrete wavelet transform to decompose the signal into wavelets at different scales and location in order to perform coding and decoding of the signal [13]. Rajeswari et al., utilized wavelet packet transform for fault diagnosis [14]. Vinod Kumar et al., has utilized a joint feature extraction of MFCC for continuous Telugu speech recognition [15]. Dinesh Sheoran et al., also discussed on spectral feature extraction techniques [16]. Lee et al., has presented prosody embeddings for speech synthesis networks [17]. Sun et al., has presented an interpretable latent variable model for prosody based on the Tacotron for text-to-speech model [18]. These techniques have been tried for various purposes and also for languages spoken in India. Miao et al., presented an Efficient TTS which is non-autoregressive architecture and mentioned the presented system outperforms when compared to Tacotron [19]. Mustaqeem et al., presented a simple light weight deep learning based system, 1-D, CNN and two stream deep CNN for Speech Emotion Recognition System in [20–22].

According to proposed PCV technique, speech is segmented into phonemes and the cepstrum variation is analyzed. The proposed feature extraction technique is unique in its own way to identify the cepstrum variation when phoneme is being analyzed. Phoneme is a sound element that differentiates the words in a language. Phonemic representation is an essential part in speech synthesis process. So, it makes sense using phoneme with cepstrum variation based feature extraction from speech. Fig. 2 shows the design of proposed speech feature extraction model. The input Tamil dataset is stored in a file and its corresponding speech dataset is created and stored separately. The creation of speech dataset is done by recording the speech in a noise free environment. In the first step, text data is analyzed for extracting the useful features. The use of intonation functions along with special symbols facilitates predicting the rise fall pattern of F0 in the sentence. In connection to the text feature extraction, features must be extracted from the speech.

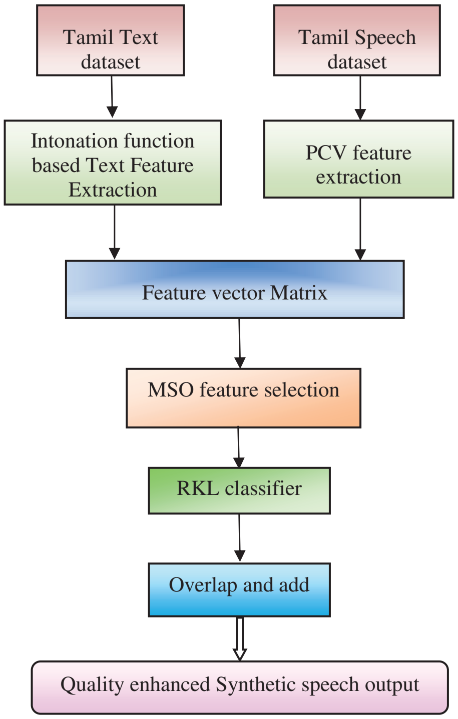


Figure 2: System design of phoneme with cepstrum variation based speech feature extraction

The research work proposes a new speech feature extraction technique which is based on the phoneme present in the speech. The speech data is stored as wav file. The speech waveform is observed in a short-term spectrum to extract the phonemic information. The PCV is the feature of interest in speech. It happens to be meaningful to know the salient traits of the cepstrum. The cepstrum is obtained by taking inverse fourier transform of the logarithmic spectrum of the speech signal. It can be realized in four different ways such as real, complex, power and phase cepstrum.

The cepstrum has the advantage in determining the pitch as it is very effectual in separating the pitch and formants whereas it is not so in the logarithmic representation of the power spectrum. To authenticate this, power cepstrum is predominantly used as feature vector to represent more useful features of speech and music. The cepstrum is obtained from the spectrum after it is subjected to transformation using the mel scale. These mel frequency cepstral coefficients are very useful in determining the pitch. It is also advantageous to use cepstrum for the purpose of pitch detection because the low frequency periodic component and formants convolve in time domain and multiply in frequency domain respectively. Another extraordinary feature of cepstrum is that it is capable of expressing convolution of two signals as an addition of their complex cepstra.

Pseudo code 2: Speech Feature Extraction-Phoneme with Cepstrum Variation

```
INPUT:          Tamil speech
OUTPUT:         Features
FRAME:          Speech Segment
TRANSFORMATIONS: {FFT, FrFT, DWT}
BEGIN
FOR each FRAME F
RETRIVE Phoneme with Cepstrum Variation P(F);
OPTIMIZE P(F) using Mosquito Spreading Optimization;
CLASSIFY P(F) using Redundant Kernel Learning;
COMPUTE TRANSFORMATIONS
INTERPRET Time t(F), Frequency f(F)
EXTRACT PCV based Transformed Features;
```

The pseudo code 2 illustrates the complete process of speech feature extraction which includes identifying and extracting the phoneme where variation is found in the cepstrum as speech features, selecting the optimal features using mosquito optimization algorithm and classifying the features using Redundant Kernel Learning algorithm.

4 Phoneme with Cepstrum Variation Based Feature Extraction

Let X be the input speech signal. It is subjected to preprocess and represented as X_k. The signal is presented in Discrete Cosine Transform as a preprocessing step for feature extraction.

$$X_K = \left(\frac{2}{N}\right) \sum_{n=0}^{\frac{1}{2}N-1} Ph(k).\cos\left(\frac{\pi.h}{2.N}(2i+1)\right) \tag{1}$$

where h–Denotes Coefficient value of transformation

After preprocessing, the signal must be normalized to obtain mean normalization for PCV feature extraction and is represented as,

$$\mu(m) = \lambda_\mu \mu(m-1) + \frac{1-\lambda_\mu}{L} \sum_{l=0}^{L-1} X(k,\ l) \tag{2}$$

This signal is presented to undergo windowing process and segmented into frames. Each frame is observed for phoneme having variation or difference in the pitch. This variation is examined with the help of cepstrum of the signal. The pitch determination is a very important step in extracting the meaningful and useful features. The phoneme in which cepstrum variation is observed is represented as,

$$Ph(k) = -ln\left\{\frac{\prod_{i=k+1}^{N} \lambda_i^{\frac{1}{N-k}}}{\frac{1}{N-k}\sum_{i=1}^{N} \lambda_i}\right\}^{(N-k)n} + M\left(\frac{1}{2} + ln(X_k)\right) - \frac{M}{k}\sum_{i=1}^{k}\left(ln\left(\lambda_i\sqrt{\frac{2}{n}}\right)\right) \tag{3}$$

where X_k–preprocessed input samples and M–average of X_K.

The advantage of extracting the speech feature using phoneme and pitch variation according to Eq. (3) is that the entire speech signal can be observed as the signal is continuous in nature. It is operated to observe the signal thoroughly without leaving any portion of the cepstrum. X_K, denotes the current phoneme in Eq. (3) and rest of the portion denotes the left and right position of the current phoneme. The intensity of pitch is determined by taking the difference between preprocessed signal and original signal. It is denoted by λ_i and described as,

$$\lambda_i = X_K - \bar{X} \tag{4}$$

where X_K is the preprocessed speech and $\overline{X}$ is the original speech.

5 Mosquito Spreading Optimization Algorithm

After extracting the features, optimal features must be selected. A new Mosquito Spreading Optimization (MSO) algorithm is proposed to select optimal features. The key idea behind the algorithm is to estimate the fittest particles as optimal features. The optimization algorithm works as follows.

The algorithm takes Feature matrix ‘T’ as an input and returns Selected Feature ‘ST’ as output. The Particles are provided as training features *f(x)*. The center position of the Particles is extracted randomly by,

$$C_{center} = ((\max(f(x)) - \min(f(x)) * Rand) + \min(f(x))) \tag{5}$$

where d–Random value (Range from 0 to 1).

The initial fitness value is extracted using,

$$Best_{fit} = ((Var_{high} - Var_{low}) * f(y)_{1,2,\ldots npar}) + Var_{low} \tag{6}$$

Where $f(y)$ – Random particles, Var_{high} – Higher limit of $f(x)$, Var_{low} – Lower limit of $f(x)$ and $npar$ – Number of feature particles.

The next step is to estimate the maximum profit of initial iteration using the objective function given by,

$$Max_{Pro} = W + \sum_{i=1}^{N} f(x)_i^2 + (\sin(2 * pi * f(x)_i)) \tag{7}$$

where W–Maximum weight of Particles.

Initialize i = 1 and iterate the loop to extract cluster formation with objective function updated by,

```
For $i$ = 1 to Number of iteration pp
If ($i \leq Max_{it\_r}$) && ($Max_{Pro}$>Accuracy)
{
        Update Particle, ‘$Particle_{weight}$’ and co-ordinates
}
```

The particle$_{\text{weight}}$ and co-ordinates are updated using,

$$Particle_{weight} = ((Particle_{Max} - Particle_{Min}) * Rand) + Particle_{Min} \tag{8}$$

$$X_{Co-ordinate}(n) = x(n-1) + \left(\left(Rand^{-\frac{1}{alpha}}\right) * \cos(Rand * 2 * pi)\right) \tag{9}$$

$$Y_{Co-ordinate}(n) = y(n-1) + \left(\left(Rand^{-\frac{1}{alpha}}\right) * \cos(Rand * 2 * pi)\right) \tag{10}$$

The distance between the particles and its corresponding weight updates is,

$$Particle_{Dist} = P(m) * (Dist * (Var_{high} - Var_{low})) \tag{11}$$

Where $P(m)$–Probability of number of Weights at each center location.

The weight of calculated distance is updated using,

$$Dist_{Update}(l) = (-1)^{Rand_{0,1}} * Dist_{Pre} * \cos\left(Dist_{Num} * \left(\frac{2}{Dist_{Num} - 1}\right) + Dist_{Pre} * \sin\left(Dist_{Num} * \left(\frac{2}{Dist_{Num} - 1}\right)\right.\right. \tag{12}$$

The maximum profit is updated using,

$$Max_{Pro\_update} = W + \sum_{i=1}^{N} C(x)_i^2 + (\sin(2 * pi * C(x)_i)) \tag{13}$$

Then set of Particles Position and Weight Position is updated by,

$$C(x) = Particle_{Population},\ Weight_{Position} \tag{14}$$

```
If ($Max_{Pro\_update} > Max_{Pro}$)
{
Update $Max_{Pro\_update}$;
Update $CK_{center}$ to New Position
}
```

The updated Profit and Particles Center continue to form Cluster till the maximum number of iterations. Finally, the Updated Particles Center is extracted as Best Fitness value output $^{'}BF^{'}$.

$$ST = T(R,\ \ BF > avg(BF)) \tag{15}$$

6 Redundant Kernel Learning Algorithm

The selected optimal features are further classified using Redundant Kernel Learning (RKL) algorithm. The optimal features are given as input to the learning algorithm and RKL applied is expressed as,

```
L = RKL (ST)
Then Signal-to-Noise ratio is computed for retrieved signal ‘L’
if SNR satisfied
{
    Return L
else
{
    Goto initialization of particles;
}
}
```

Pseudo code 3: Redundant Kernel Learning

```
INPUT: Updated training feature set ST
Testing feature V
OUTPUT: Classified Label CL
BEGIN
INITIALIZE Probability array p
            $p = P(q_i = s_i)| \ i = 1 \text{ to } n$
FOR all features ST
EXTRACT labels L(ST)
COMPUTE probability to estimate training feature P(ST)
COMPUTE probability to estimate testing feature P(V)
IS P(ST)>P(V)
CL = L(P(V))
END
```

The pseudo code 3 illustrates how the learning algorithm is applied to classify the features and label them. The RKL learning algorithm is a special kind of machine learning algorithm which takes the training and testing data set represented in the form of matrix as its input and returns classified label after learning. The training set comprises the optimal features selected using MSO denoted by ST_r, the testing set denoted by V and initialize label L.

```
    Initialize the Probability array,
    $p = P(q_i = s_i)$ where, s–State of training set for $i$ = 1, 2… N
        N–Size of Training set ‘ST’
        q–Fixed state sequence for the length of ‘V’
    for ($i$ = 1 to Row_size (V))
    for (j = 1 to Column_size(V))
$s_i = STr(i, j)$
$d_i = \sqrt{(s_i - V_i)^2 + (s_j - V_j)^2}$
    $q_{i,j} = L(d_i)$; //Extract Corresponding labels of Training set.
$\pi_i = \frac{\sum_{k=1}^{m} s_i(q_{i,j}(d))}{m}$
$\sigma_T = \sqrt{\frac{1}{m}\sum_{i=1}^{N} (s_i(q_{i,j}(F)) - \pi_i)^2}$
$\sigma_V = \sqrt{\frac{1}{n}\sum_{i=1}^{N} (V_i(q_{i,j}(d)) - \pi_i)^2}$
where ‘m’–length of $s_i$ and n’–length of $V_i$
    Probability of estimating the training feature set
$P(STr|\pi_i) = \left(2\prod \sigma_T^2\right)^{\frac{-N}{2}} * \mathrm{e}^{\left\{\left(\frac{-1}{2<_T^2}\right)\|STr - \pi_i\|^2\right\}}$
where N-size of training set ‘$STr$’
    Probability of estimating the testing set
$P(V) = \left(2\prod \sigma_D^2\right)^{\frac{-M}{2}} * \mathrm{e}^{\left\{\left(\frac{-1}{2\sigma_D^2}\right)\|V_i - \pi_i\|^2\right\}}$
where M–size of testing set ‘V’
          Check the condition for verifying the features, if condition satisfied, classify using label
          if ($P(STr|\pi_i) > P$ ($V$))
                    $CL_i = L$ ($P(V)$);
          end if
      end ‘j’ loop
    end ‘$i$’ loop
```

A probability array p is also initialized. Si stores the extracted attributes of training set. $q_{i,j}$ holds the corresponding labels extracted from the training set. The probability of training features and testing features are estimated. If the probability of training feature is greater than the testing feature set, then classification is performed using neural network model and labeled. The optimal features obtained are used for learning purpose with the help of learning algorithm which is fed to the neural network model for further classification. Then, SNR values are computed for the retrieved signal and labeled according to the classification.

The results are improved by 14% using phoneme with cepstrum variation based feature extraction when compared to transformation based feature extraction techniques.

7 Simulation Results and Discussion

The simulation setup is created using MATLAB and different Tamil speech signal segments are chosen from the dataset. The signals are sampled at 8 KHz and distorted using white noise and street noise. The length of the test speech data segment is 180.6 s including the silence period. During the experimental analysis, 85% of the speech segments are classified as voiced speech.

The Fig. 3 shows the input speech signal and the signal after normalization is applied. The usual way of normalizing the speech is to multiply the signal by a factor of 1/max (signal).

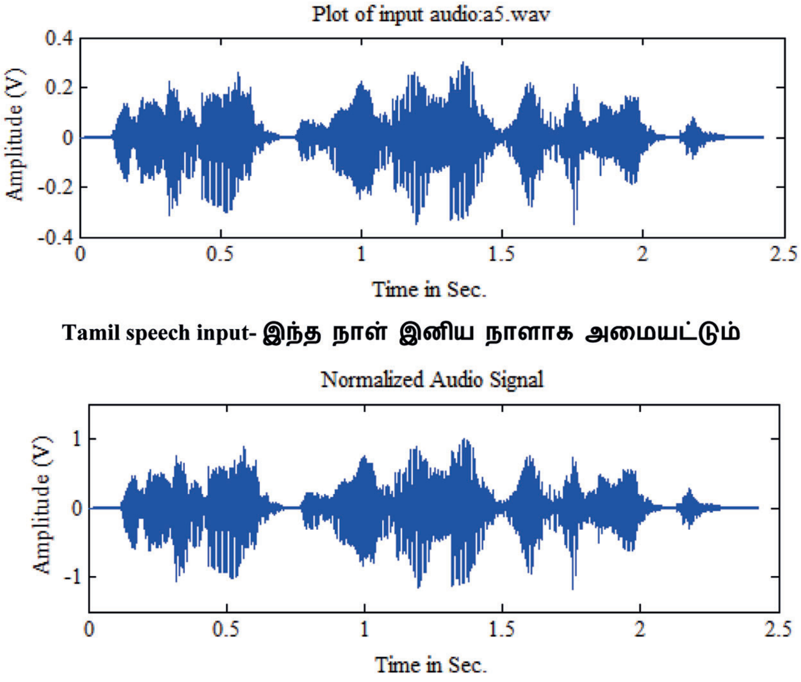


Figure 3: Input speech signal and normalized signal

Figs. 4–6 shows the noise filtered signal, its initial spectrum for feature extraction and power spectrum of the speech respectively.

7.1 Speech Signal Using PCV Based Features After Applying Transformation

Fig. 7a shows the FFT spectrum of the processed speech, its pitch track, its peak point representation and the quality improved output. It is evident from the peak points that the rise and fall marked in red triangle and circle points indicate the places where pitch is to be modified. The overall pitch range can also be referred clearly with the help of pictorial representation of the pitch track of the signal. Similarly, the FrFT and DWT spectrum and its corresponding information are obtained with the help of Figs. 7b and 7c respectively.

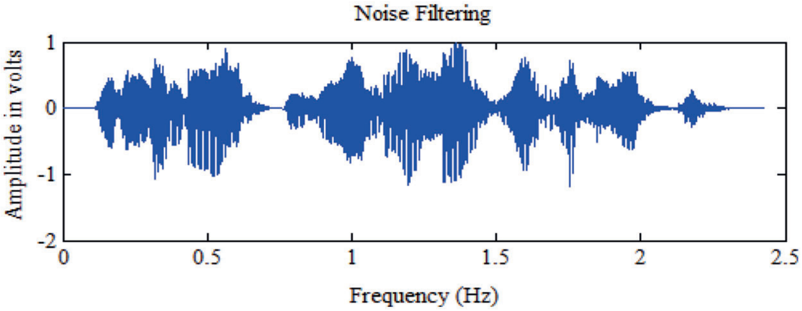


Figure 4: Noise filtered Signal

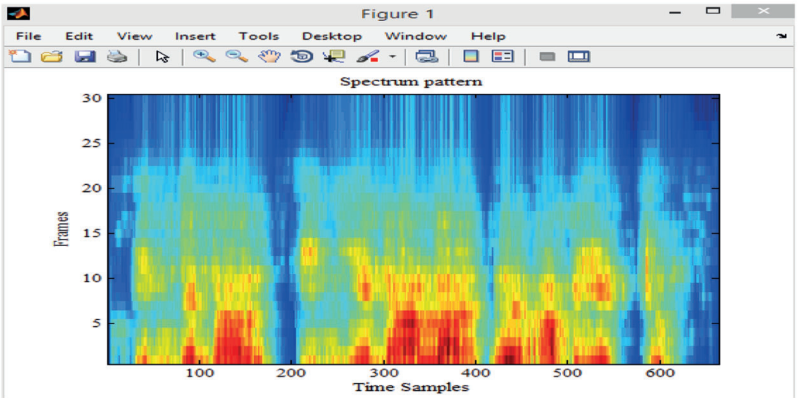


Figure 5: Spectrum pattern of the input sentence

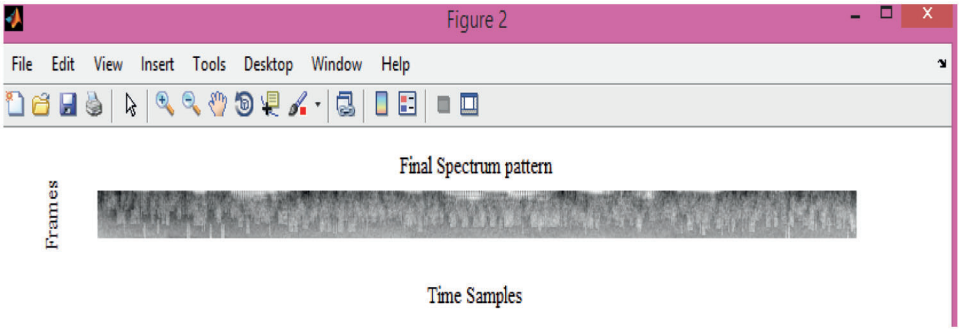


Figure 6: Power spectrum pattern of the input sentence

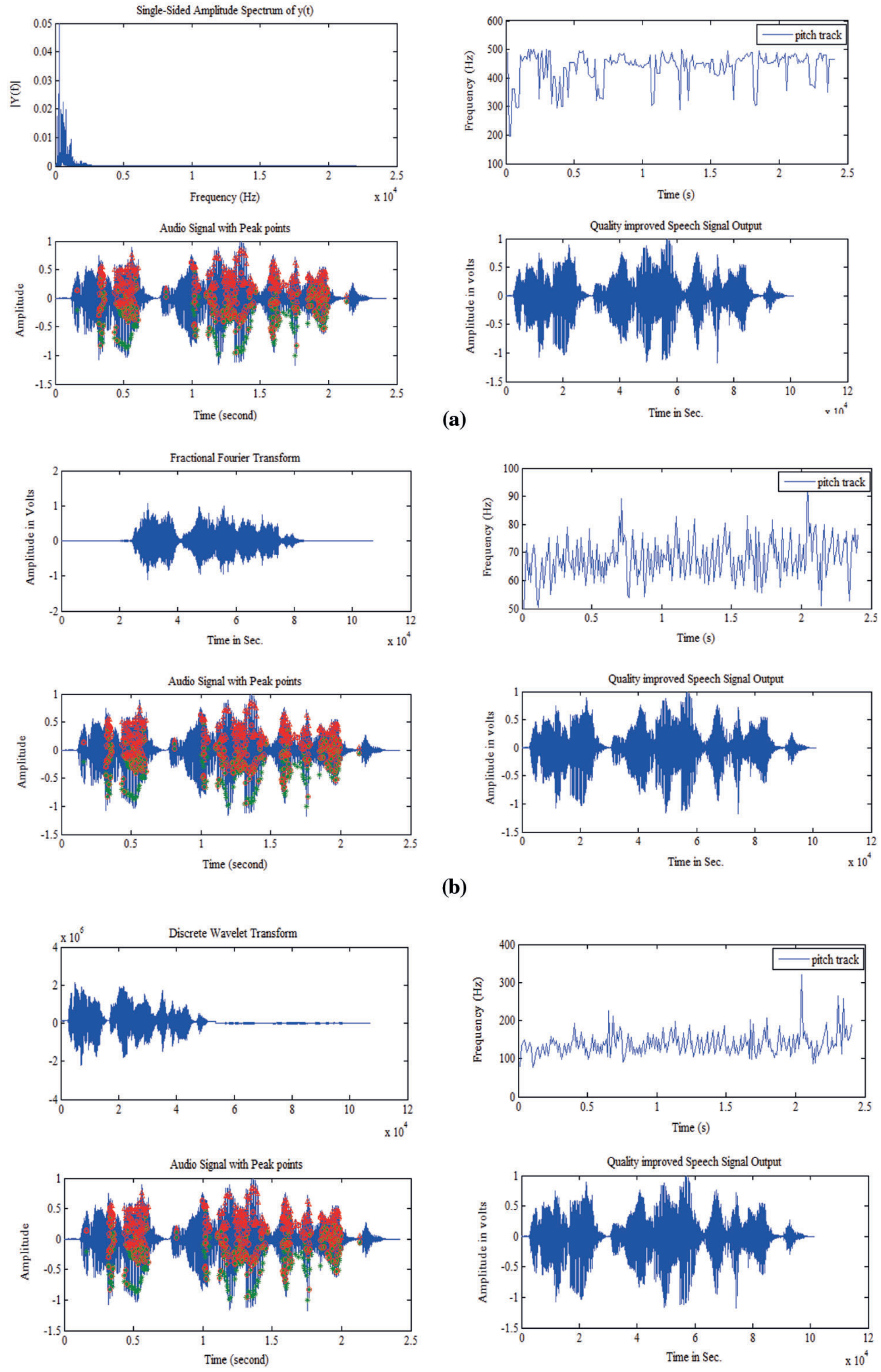
7.2 Objective Measures

The proposed PCV based feature extraction technique is evaluated using the objective measures such as Signal-to-Noise ratio (SNR), and Normalized Correlation coefficient (NCC).

Signal-to-Noise ratio (SNR): The Signal-to-Noise is one of the simplest and promising measures to evaluate the quality of the speech in terms of intelligibility. It is measured as the ratio of signal power to the noise power and represented by

$$SNR = 10 \log 10 \frac{Psignal}{Pnoise} \tag{16}$$

where P_{signal} is the power of the signal and P_{noise} is the noise power. SNR is always expected to be better associated with the Mean Opinion Score (MOS). It is also simple to compute and has been used predominantly to qualify the enhanced speech.



(a)

(b)

(c)

Figure 7: (a) FFT, (b) FrFT and (c) DWT with pitch track, peak points and quality enhanced speech output

Normalized Correlation Coefficient (NCC): Correlation provides the similarity between two signals. Normalized correlation coefficient (NCC) ranges between 1 and −1. The upper bound of NCC is 1 and lower bound is −1. It is good to have larger NCC values, because greater the NCC values, more similar will be the signals. It is calculated as follows

$$Correlation\ coefficient = \sum_{n=0}^{N-1} x(n)y(n) \tag{17}$$

$$NCC = \frac{\sum_{n=0}^{N-1} x(n)y(n)}{\sqrt{\sum_{n=0}^{N-1} x^2(n) \sum_{n=0}^{N-1} F^2(n)}} \tag{18}$$

7.3 Subjective Measures

Mean Opinion Score (MOS) is one of the most extensively used subjective measures to assess the quality of enhanced speech with respect to intelligibility and naturalness. It is a kind of listening test conducted using properly designed experimental setup. The listeners are called as subjects. The MOS values are scaled between 1 and 5. The MOS values of 1 indicates that the quality of speech is very poor, 2 indicates poor, 3 indicates fair, 4 indicates good and 5 indicates the quality is excellent. The choice of subjects can be native language speakers, non-native language speakers, age between 25 and 35 or age above 40.

SNR Results: Tab. 3 shows the comparison of SNR values obtained while prosody modification is done using information provided by FFT based MFCC feature, and PCV based feature. The comparison is done for the various types of intonation functions against various types of the sentences. It is evident from the table that PCV based feature provides useful information for pitch modification compared to FFT based MFCC features. The value of SNR for PCV based feature is approximately two-fold increase compared to FFT based MFCC feature extraction technique.

Table 3: SNR results for FFT based MFCC and PCV feature extraction

Tamil sentence and its intonation function	Type of sentence	SNR value	
		FFT based MFCC	PCV based feature
உங்கள் சொந்த ஊர் எது?-?–Grammatical function	Interrogative sentence	27.035	53.316
இந்த நாள் இனிய நாளாக அமையட்டும்.-focus function	others	26.246	52.527
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	27.156	53.431
நீ வரும் வரை, நான் காத்திருப்பேன்.-discourse function	Declarative sentence	26.314	52.595

When the proposed intonation function based text feature extraction and PCV based speech feature extraction techniques are examined to identify the sentence type using special symbols present in the sentence and intonation function, the SNR value for the exclamatory sentence is higher than the other types of sentences. Tabs. 4 and 5 shows the comparison of SNR values obtained while prosody modification is done using information provided by FrFT based MFCC feature Vs PCV based speech feature, and DWT based MFCC Vs PCV based speech feature extraction respectively.

Table 4: SNR results for FrFT based MFCC and PCV feature extraction

Tamil sentence and its intonation function	Type of sentence	SNR value	
		FrFT based MFCC	PCV based feature
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	28.291	55.811
இந்த நாள் இனிய நாளாக அமையட்டும்.-focus function	others	27.502	55.022
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	28.412	55.932
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	27.570	55.070

Table 5: SNR results for DWT based MFCC and PCV feature extraction

Tamil sentence and its intonation function	Type of sentence	SNR value	
		DWT based MFCC	PCV based feature
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	29.392	58.304
இந்த நாள் இனிய நாளாக அமையட்டும்.-focus function	Others	28.603	57.515
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	29.513	58.425
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	28.671	57.583

Figs. 8–10 shows the comparison of SNR values obtained for FFT, FrFT and DWT Vs PCV respectively.

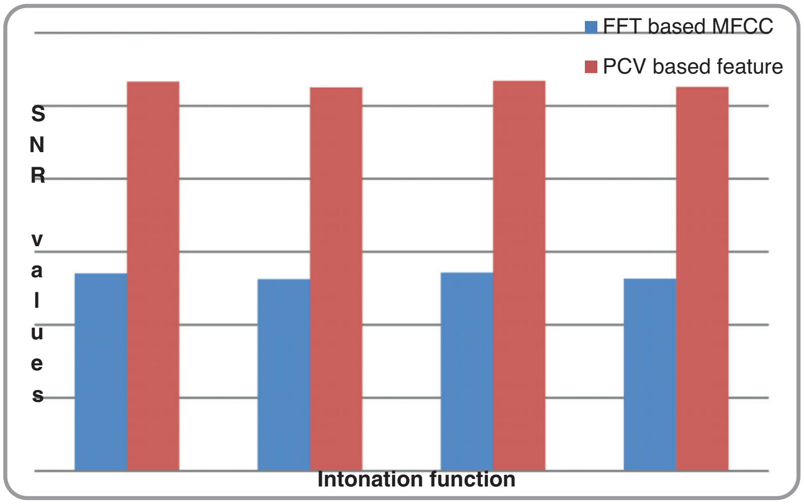


Figure 8: Comparison chart of SNR results for FFT based MFCC and PCV based optimized feature

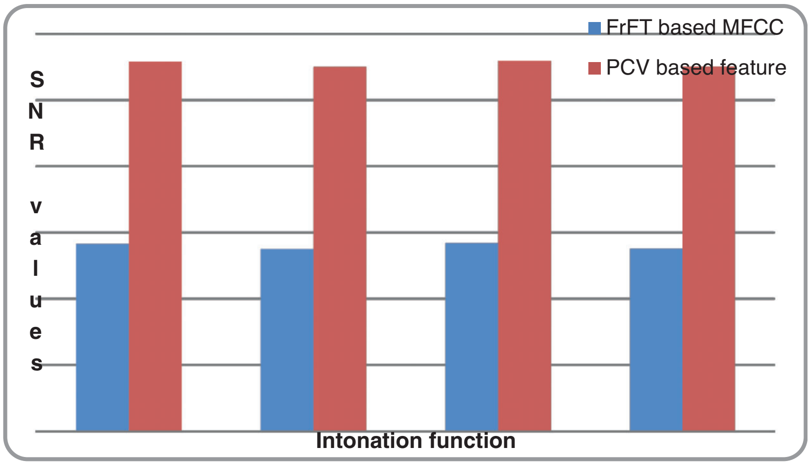


Figure 9: Comparison chart of SNR results for FrFT based MFCC and PCV based optimized feature

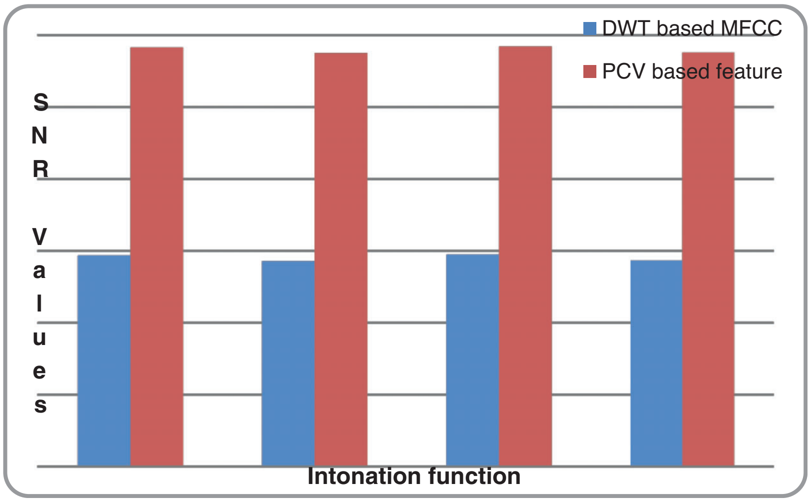


Figure 10: Comparison chart of SNR results for DWT based MFCC and PCV basedoptimized feature

NCC Results: Tab. 6 shows the comparison of NCC values between unprocessed signal and processed signal. The similarity check is done with reference to the natural speech produced by proficient native Tamil speaker. The unprocessed signal is the synthetic speech without being subjected to any quality improvement related process such as noise filtering, feature extraction and selection to incorporate prosody, and special production techniques. The processed signal is the synthetic speech subjected to quality enhancement related processes. The NCC values are comparatively high for processed input signal. The NCC values are also observed for Transformation features, MFCC features, Transformation based MFCC and PCV based features for each of the sentence. The NCC is an important objective measure always proportionate to the MOS subjective test results.

Table 6: NCC values for unprocessed and processed Tamil sentences

Tamil sentence and its intonation function	Type of sentence	NCC value	
		Unprocessed signal	Processed signal
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	0.58	0.93
இந்த நாள் இனிய நாளாக அமையட்டும்.-focus function	others	0.61	0.95
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	0.75	0.98
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	0.55	0.85

Tab. 7 shows the NCC values obtained for the input sentences are subjected to feature extraction through DWT transform, MFCC, Transform based MFCC and proposed PCV based technique.

Table 7: NCC values for Transform, MFCC, TR-MFCC and PCV feature extraction

Tamil sentence and its intonation function	Type of sentence	NCC value in %			
		Transform	MFCC	Transform based MFCC	PCV feature extraction
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	87.980	89.061	91.990	95.205
இந்த நாள் இனிய நாளாக அமையட்டும்.-focus function	others	87.432	87.932	89.546	92.761
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	88.032	89.912	92.032	95.247
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	87.652	88.921	91.205	94.410

Figs. 11 and 12 shows the comparison of NCC values of speech signal subjected to existing speech feature extraction and proposed PCV based speech feature extraction technique.

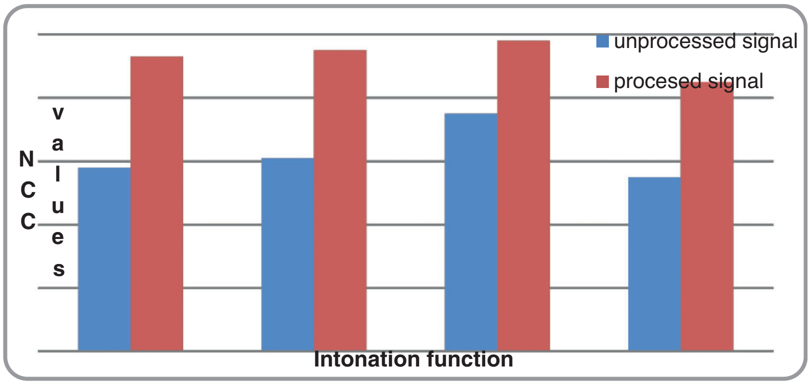


Figure 11: Comparison chart of NCC values for unprocessed and processed signal

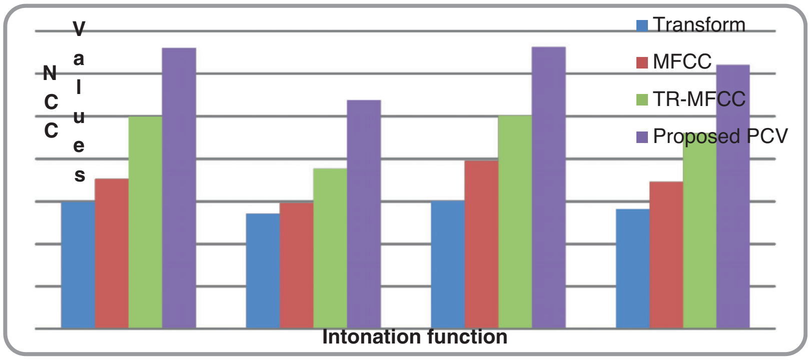


Figure 12: Comparison chart of NCC values transform, MFCC, transform based MFCC and proposed PCV based feature

MOS Results: Tab. 8 shows the MOS results with respect to intelligibility and naturalness of the enhanced speech using the proposed technique of PCV based optimal speech feature extraction in order to incorporate appropriate intonation component in Tamil TTS.

Table 8: MOS values obtained from native Tamil speakers

Tamil sentence	Type of sentence	MOS value obtained from native Tamil speakers with respect to naturalness of the synthesized speech				
		S1	S2	S3	S4	S5
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	4.08	4.09	3.86	3.93	4.2
இந்த நாள் இனிய நாளாக அமையட்டும்.- focus function	others	3.68	3.75	3.69	3.72	3.99
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	3.9	4.0	4.2	4.12	4.25
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	4.1	4.09	3.86	3.93	4.2

It is observed from the Tabs. 8 and 9 that the average MOS in terms of naturalness perceived by Tamil speakers are 4.08 for interrogative sentence, 3.76 for others category, 4.09 for exclamatory sentence and 4.036 for declarative sentence. The average MOS scores obtained from non-native Tamil speakers for the set of sentences is 4.01, 3.83, 4.08 and 4.00 respectively. The MOS values show good score for exclamatory sentences and are provided by both native and non-native Tamil speakers. This result validates the proposed technique of speech feature extraction in which extracting phoneme with cepstrum variation is a feature of interest.

Table 9: MOS values obtained from non-native Tamil speakers

Tamil sentence	Type of sentence	MOS value obtained from non-native Tamil speakers with respect to naturalness of the synthesized speech				
		S1	S2	S3	S4	S5
உங்கள் சொந்த ஊர் எது?-Grammatical function	Interrogative sentence	3.99	3.95	4.09	3.94	4.10
இந்த நாள் இனிய நாளாக அமையட்டும்.- focus function	others	3.78	3.59	3.93	3.98	3.88
ஓ! இவ்வளவு இந்தியர்கள் எம்மை ஆதரிக்கிறார்களே!-indexical function	Exclamatory sentence	3.85	39.9	4.16	4.21	4.22
நீ வரும் வரை, நான் காத்திருப்பேன்-discourse function	Declarative sentence	3.89	3.95	3.96	4.00	4.20

8 Conclusion

This paper highlights the enhanced quality of speech in Tamil TTS. The work has proposed a completely new text and speech processing techniques. Special symbols present in the text enables extracting the intonation information, a prosodic component as the meaningful text feature. Speech feature extraction using phoneme with cepstrum variation as the feature choice is presented. The PCV based features are extracted and further selection is done to extract optimal features using Mosquito spreading optimization algorithm. A RKL learning classifier is used to classify the labeled outputs. The result is used to incorporate necessary prosody modification when the speech is synthesized. The merit of the proposed system is it eradicates the need to study the actual human articulatory mechanism which is extensive and tiresome process. The performance of the proposed approach is compared against the existing feature extraction techniques. The comparison results of objective measures such as SNR values, NCC values and subjective measure MOS shows the improvement in the quality of speech, in accordance with the proposed approach. From the results, it is evident that the quality of synthetic speech is enhanced by 14% when the proposed PCV based speech feature extraction is used rather than the transformation based MFCC.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. T. Makwana and D. C. Vegda. "Survey: Natural language parsing for Indian languages," arXiv preprint arXiv:1501.07005, pp. 1–9, 2015.

[2] B. Sudhakar and R. Bensraj, "Enhanced evaluation of sentiment analysis for tamil text-to-speech synthesis using hidden semi-markov model," *Communication on Applied Electronics*, vol. 3, no. 6, pp. 13–16, 2015.

[3] V. J. Pooja, B. M. Smitha, K. K. Pratik and R. R. Desmukh, "A comparative study between MFCC and DWT feature extraction technique," *International Journal of Engineering Research & Technology*, vol. 3, no. 1, pp. 3124–3127, 2014.

[4] K. C. Rajeswari and P. UmaMaheswari, "Feature extraction and analysis of speech quality for tamil text-to-speech synthesis system using fast Fourier transform," *Australian Journal of Basic and Applied Sciences*, vol. 9, no. 35, pp. 349–356, 2015.

[5] N. Shreya and G. Divya, "Speech feature extraction techniques: A review," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 3, pp. 107–114, 2015.

[6] A. Apte and H. Gado, "Tamil character recognition using structural features," 2010.

[7] V. Ramu Reddy and K. Sreenivasa Rao, "Two-stage intonation modeling using feed forward neural networks for syllable based text-to-speech synthesis," *Computer Speech and Language*, vol. 25, no. 5, pp. 1105–1126, 2013.

[8] A. R. Beigi and A. E. Rasekh, "Paragraph intonation patterns among Persian speakers' English interlanguage: The diversity of paratones in focus," *Covenant Journal of Language Studies (CJLS)*, vol. 3, no. 1, pp. 1–17, 2015.

[9] J. Romportl and J. Kala, "Prosody modeling in Czech text-to-speech synthesis," in *Proc. Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany, pp. 200–205, 2007.

[10] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *International Journal of Computer Applications in Engineering, Technology and Sciences*, vol. 2, no. 2, pp. 412–418, 2013.

[11] M. Shafi, M. S. Khan, N. A. Sattar, M. Rizwan, A. A. Baba *et al.,* "Transform based speech enhancement using DCT based MMSE filter and its comparison with DFT filter," *Journal of Space Technology*, vol. 1, no. 1, pp. 47–52, 2012.

[12] P. K. Kurzekar, R. R. Deshmukh, V. B. Waghmare and P. P. Shrishrimal, "A comparative study of feature extraction techniques for speech recognition system," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, no. 12, pp. 18006–18016, 2014.

[13] V. K. Yadav, A. Jain and L. Bhargav, "Analysis and comparison of audio compression using discrete wavelet transform," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 1, pp. 310–313, 2015.

[14] C. Rajeswari, B. Sathiyabhama, S. Devendiran and K. Manivannan, "Bearing fault diagnosis using wavelet packet transform, hybrid PSO and support vector machine," *Procedia Engineering*, vol. 97, no. 1, pp. 1772–1783, 2014.

[15] V. K. Sharma and A. P. Kumar, "Continuous telugu speech recognition by joint feature extraction of MFCC, MODGDF and DWPD techniques by PNN classifier," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 865–872, 2018.

[16] D. Sheoran, P. Sangwan and M. Khanna, "Spectral feature extraction techniques for speech recognition," *International Journal of Multidisciplinary Research and Development*, vol. 4, no. 6, pp. 33–38, 2017.

[17] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Brighton, United Kingdom, pp. 5911–5915, 2019.

[18] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen *et al.,* "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. ICASSP*, Barcelona, Spain, pp. 6264–6268, 2020.

[19] C. Miao, S. Liang, Z. Liu, M. Chen, J. Ma *et al.,* "Efficient TTS: An efficient and high-quality text-to-speech architecture," in *ICML*, pp. 7700–7709, 2021. https://icml.cc/Conferences/2021.

[20] Mustaqeem and S. Kwon, "Att-net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, pp. 107101, 2021.

[21] Mustaqeem and S. Kwon, "1d-CNN: Speech emotion recognition system using a stacked network with dilated CNN features," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 4039–4059, 2021.

[22] Mustaqeem and S. Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems*, 2021. [Online]. Available: https://doi.org/10.1002/int.22505.