ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Context Based Segmentation and Spectral Mismatch Reduction for More Naturalness with Application to Text to Speech (TTS) for Marathi Language

Mrs Smita Kawachale[1], Dr. J. S. Chitode[2]

Research Scholar, Electronics Department, Bharati Vidyapeeth College of Engineering, Pune, India[1]

Honorary Professor, Electronics Department, Bharati Vidyapeeth College of Engineering, Pune, India[2]

Abstract: The field of text to speech (TTS) synthesis has been rapidly developing with widespread applications. There is a great demand for text to speech synthesis for Indian languages. TTS in English and world's most used languages are been developed already. The proposed work is for Text to Speech conversion for Marathi language. This TTS is capable of speaking Marathi text. It is using 'Hybrid Syllabic Approach' where it forms and speaks new words from the syllables derived from the existing words in the database. Syllabic based speech synthesis is based on Consonant Vowel (CV) structure rules. An optimized soft cutting (segmentation) approach is followed for more naturalness and improved context based database.

The proposed work focuses on improving naturalness of TTS using context based segmentation. Context based segmentation is based on syllable position (I-Initial, M-Medium, F-Final). The proposed work focuses on position dependent (I/M/F) speech synthesis. Concatenation of position dependent syllable may result in less spectral mismatch (concatenation cost) and give more natural sounding audio output. By carrying out this spectral analysis it is possible to improve the naturalness and overall performance of TTS. Spectral mismatch reduction is carried out with different Time and Frequency domain parameters. The performance of proposed method is evaluated using Subjective and Objective validation methods.

Keywords: Text to Speech System, Spectral Smoothing, Concatenative TTS, Speech Synthesizer.

I. INTRODUCTION

It is extremely tough to make a machine which sounds identical to human. Hence the best text to speech (TTS) algorithm ever made sounds robotic, until and unless human speech itself is involved in it. But it is not possible to create a database of each and every word possible in any language. Syllable based Concatenative Speech Synthesis (CSS) leads to formation of new words from existing words in data base. The most important qualities of a speech synthesis system are naturalness and intelligibility.

Two basic methods of speech synthesis are, (1) Rule based synthesis: Rule based speech synthesis uses rules of particular language to generate the synthetic speech. (2) Dictionary based synthesis: Dictionary based speech synthesis uses most commonly used words in the audio database. Rule based synthesis has the drawback of reduced naturalness of synthetic speech and Dictionary based synthesis has large database size as each word needs to be stored. But syllable based speech synthesizer generates more number of words based on very small database. Different syllables can form new words. Hence original database is not large. Study of various types of synthesizers shows that among all types of speech units, syllable results in more naturalness. An intelligible text- to-speech program allows people with visual impairments or reading disabilities to listen to written material on a home computer [1][2].

The key objective of proposed work is to design a system that develops syllables automatically from different words. Manual segmentation being very time consuming, supervised and unsupervised methods of neural network can be used for the development of proper syllable formation. The importance of neural network in syllabification is:

1. Its ability to generalize and capture the relationship between input and output pattern pairs.
2. Its ability to predict, after an appropriate learning phase, even those patterns not presented before.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

3. Its ability to tolerate certain amount of fault in input.

Vowel detection can be done by calculating energy of sound file. Vowels have more energy as compare to consonants and hence syllables can be cut very easily by making use of this property. Neural Network algorithms are used in proposed work for carrying out proper segmentation of speech. For comparing the accuracy some non neural approaches along-with neural networks are discussed. Neural approaches are providing more than 90% accuracy in syllable segmentation while the accuracy of non-neural approaches is limited to less than 70%. Concatenative Speech Synthesis is widely used due to its naturalness and less signal processing requirement. But it has problems like requirement of large database and resulting spectral mismatch in output speech. In concatenative TTS position of syllable plays very important role while carrying out segmentation.

If proper position syllable is used while forming new words from existing syllables, resulting spectral mismatch is less. If position of syllable is not considered during concatenation of speech units, resulting synthesis end up in more concatenation cost. This work presents different techniques like PSD, Wavelet and DTW to find spectral mismatch in concatenated segments. If syllable position is considered while forming new word then it is called Properly Concatenated (PC) word and if new word is formed from available syllables without considering their position it is called Improperly Concatenated (IC) word. It is observed that formant plots for original and properly concatenated words are very similar for all formants while for improper concatenation extra peaks are observed in all formants. These extra peaks can be considered as estimation for spectral mismatch. With direct formant modification one can overcome spectral mismatch and smooth some of the frames which helps to reduce glitch type of sound at concatenation point. Wavelet based audio results shows more naturalness compare to other methods. In proposed work the discontinuities at the cutting point are smoothed by changing the spectral characteristics before and after the cutting point so that the spectral mismatch is equally distributed over the number of adjacent frames. [3]

II. MAIN WORK

Energy calculation is the basic requirement of syllable cutting. It is done to identify vowels in the words. Normally each syllable contains at least one vowel. Syllables are cut after vowels are identified. Consider the breaking of following word into syllables. The segmentation of syllables is relatively easy.

E.g. : (vaarkari) वारकरी =(vaar) वार + (kari)करी

here (vaar) वार and (kari) करी are two syllables.

Main Block Diagram of system is shown below.

Block Diagram

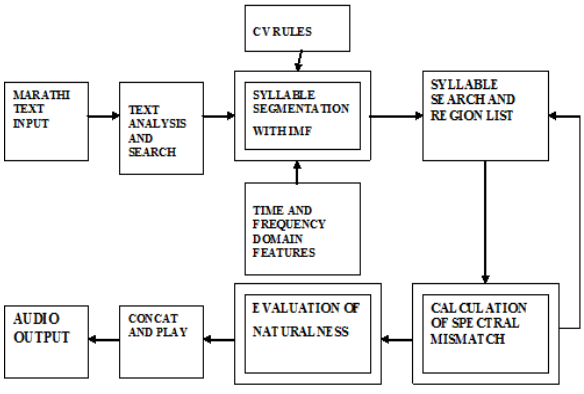


Figure 1: General framework of proposed work.

Fig.1 General framework of proposed work.

Energy of word is calculated by accessing samples of data in wave file. Wave file data is accessed sample by sample and its energy is calculated. To make samples processing easy, samples are grouped in frames. Energies of frames are plotted against frame number.

A. Calculation of Energy

1): Definition of Energy

Normally the energy of the signal is defined as

$$E = \sum_{m=0}^{m=\infty} (x(m) * x(m)) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where m varies from minus infinity to plus infinity.

2): Actual Formula Used

The above formula has little meaning for speech since it gives little information about time dependent properties of the speech signal. So short time energy at sample is defined as

$$E = \sum_{m=n-N+1}^{n} (x(m) * x(m)) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Where m varies from n-N+1 to n.

Here N is total number of samples in one frame and n is a sample number. But difficulty with the above formula is that it is very sensitive to large signal levels (since they enter in the computation as a square), thereby emphasizing large

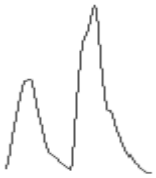
ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

sample-to-sample variation in x (m). Average magnitude function for calculating energy is given as

$$E = \sum_{m=n-N+1}^{n} |x(m)| \qquad \ldots\ldots\ldots\ldots (3)$$

Where m varies from n-N+1 to n.
This function is called average magnitude but here it is called as energy only. There should be exactly one vowel in each syllable for proper breaking of words into syllables; therefore to separate syllables from a word it is required to separate vowels. Some parameter of speech is required which can clearly differentiate between consonant and vowel. Energy is one such parameter. Energy waveform for the word मंत्री is shown below in figure 2.



मंत्री

Fig. 2 Energy waveforms of मंत्री

The CV structure of the word is CVCCCV. It consists of two vowels. So there should be two peaks in the energy waveform of this word. Thus energy graph of a word proves to be a useful tool to separate out vowels in the word leading to separation of syllables of the word.
Short Segments of speech signal is nothing but 10ms duration of speech. These short segments are called frames. Energy of every frame (E) when plotted against time produces new time dependent sequence, which can serve as a representation of the speech signal. [5]

B. Location of Vowels

Energy plot of the word is segmented in such a way that each broken part will contain one vowel. Single point must be located on the energy waveform, which will separate two vowels. Therefore, when the position of the minima between two peaks is located accurately two vowels are separated and hence the two syllables. Energy plot of Z§Va is shown below in figure 3. If minima location is calculated (pointed by arrow) through the program and wave file is played from start up to the location of minima then syllable ZZ² can be played. If wave file is played from minima to end then Va will be played.

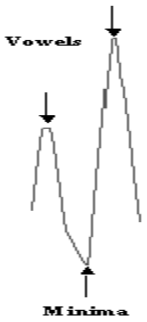


Fig. 3 Proper minima detection for word नंतर

For proper detection of minima it is required that Energy waveform should be smooth. This will prevent false interpretation of small rise and falls on the waveform as peak and minima respectively. For smoothening the waveform a technique called 'Moving Average' is used where 12 values are taken starting from the first element of the energy array, its average is calculated and stored it as the first element in another array. Then next 12 values from the second element onwards are taken, its average is calculated and preceded in similar fashion. This can be modeled as follows:

$$Smooth\ (i) = \{\textstyle\sum_{i}^{i+11}(energy\ (i)/12)\} \ldots\ldots\ldots (4)$$

Instead of 12 one can use any other value as per the smoothness required. Figure 4 below shows smoothening of energy graph of नंतर before and after averaging.

नंतर Before Averaging After Averaging

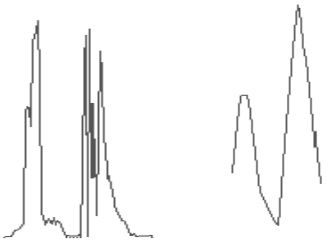


Fig. 4 Smoothing of energy waveforms

C. Results

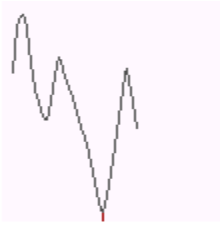
Vowels are shown below for three syllable word.

1): Three vowels word

Figure 5 below shows example of three vowel word. The three peaks clearly indicate the location of three vowels in word. As energy of vowels is large, for every vowel there is presence of one peak.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013



ज्वलंत

Fig.5 Example of three vowel words

Fig. 6 shows the block diagram of syllable cutting algorithm. Energy of the word is calculated from its audio file. Vowels are then located using the energy plot. The syllables are located using the CV structure breaking rules. From and to durations of the syllables forms the regions list.

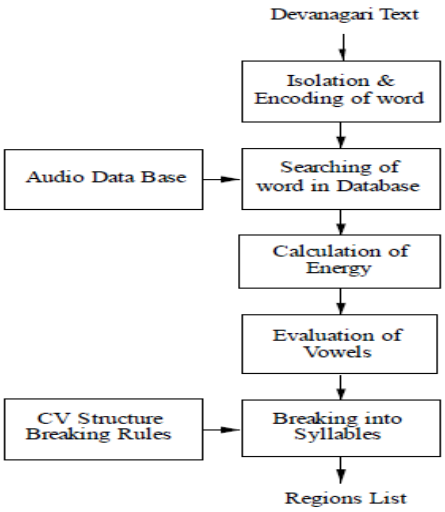


Fig.6 Block Diagram of Syllable Cutting Algorithm

2): Isolation and Encoding of Words

Although Indian standard codes for Information Interchange (ISCII) are available for development of Devanagari characters, different codes are assigned to each character. Separate ranges of codes are assigned for consonants and vowels. The entire text is converted into a string of code values shown in table 1.

TABLE 1
CODES

Type	Codes
Full Consonants	72 to 105
Half Consonants	145 to 178
Vowels	65 to 71 and 117 to 139

Each word is represented as a string of code values separated by "|". For example the word भारतीय will get converted into the code string 98|116|97|86|120|169|.

3): Handling of Shortcut Keys

There are certain shortcut keys to type more than one character using a single keystroke. For example क्व corresponds to single ASCII value 186, ध्द corresponds to ASCII value 213. Such characters are separated and assigned appropriate codes. So the code 186 is replaced by code for क (138) and code for व (86).

4): Positioning of Characters

The next stage in the text encoding is inserting the rafars, kannas, anuswer and matras in proper position. The pronunciation of refars is before the character on which it is present, whereas the pronunciation of anuswer is after the character on which it is present. For example for the word अर्विंद the pronunciation of rafars is immediately after अ while pronunciation of anuswar is after व.

5): Handling of Anuswar

The pronunciation of anuswar in Devanagari depends on the character that follows it. So when anuswar occurs, it is replaced by an appropriate half consonant depending on text character in the word. However, when anuswar is the last character of the word it has no nasal sound.

6): Handling of Special Characters

If the character is either ृ or ु then the previous consonant is made half while the character is replaced by appropriate full consonant. Some examples are illustrated below

Syllable वृ is replace by व and रु

Syllable प्र is replace by प् and र

Syllable ह्या is replace by ह् and या

7): Database Searching

Two databases are maintained: 1) Audio database, that stores the audio files and 2) Textual database that stores the text files corresponding to audio files in the audio database. The textual database is required to search the index of the required word in the audio database.

8): Audio Database

WAV file format is used for recording sound files. WAV files are probably the simplest of the common formats for storing audio samples. Unlike MPEG and other compressed formats, WAV files store samples in the raw form where no pre-processing is required other than the formatting of data. The maximum bandwidth that human speech can attain is 3.5 KHz; hence the 11 KHz, 8 bit, mono format used is sufficient for analysis of speech.

9): Calculation of Energy

For proper vowel identification energy is calculated. It is necessary to consider CV structure breaking rules and

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

periodograms. For syllable cutting the energy waveform of the word has to be as smooth as possible. For smoothing moving average technique can be implemented. The length of moving average window is 12. Generally moving window of 8 to 12 size gives good results.

10): Evaluation of Vowels

Vowel evaluation is based on checking the peaks and cutting at minima of energy. Detection of vowel can be done by detecting more or less the consonant part, which follows the vowel that is minima of energy. For fixing minimas the successive difference method is used. In this method successive difference of two adjacent samples is found. The point where the successive difference changes from positive to negative is nothing but the corresponding minima in between two vowels or peaks. Fig. 7 shows the energy plot, corresponding minimas of two words.

Energy-Peaks

Fig 7 Vowel Detection

11): CV Structure Formation

CV structures are formed based on the rules mentioned below:

1. A full consonant is assigned structure “CV”
2. A half consonant is assigned structure “C”
3. A full vowel is assigned structure “V”
4. Anuswar and rafar are considered as half consonant and assigned “C” and similar other few rules of CV structures are derived from empirical study of Marathi language and implemented in the TTS.

12): Consonant Vowel (CV) Structure Breaking

After forming the CV structure for the word the text analyzer finds the CV structures after breaking the word. There are finite numbers of distinct CV structures for a given language. They were identified by analyzing over 5000 words. The CV structures formed are split according to rules derived empirically based on the structure of Devnagri script. The database stores the list of distinct CV structures for Marathi language along with its corresponding structure after breaking. Fig. 8 shows how the syllables are cut for the word नंतर (CVCCVC). It can be cut as CVC+CVC. Hence the word will have syllables of नन् (CVC) and तर (CVC).

Fig 8 Syllable Cutting

13): Syllable Search

The syllable-searching module is identical to the word-searching module. The only difference being that syllables are searched in the textural database instead of entire words. Syllables are not recorded separately; rather they are soft cut from words stored in the audio database.

14): Concatenation

The output from the previous modules is a file containing from and to location of the required words and/or syllables. The concatenation module reads the filename and from and to locations from the file, opens the required sound file and plays the required words and/or syllables. A delay is added after each word so as to make the speech clearly audible.

15): Results and Discussion

This algorithm is tested for large number of words. It breaks the structures of almost all the words satisfactorily.
Following figure shows one example.
The word हत्ती

TABLE 2: हत्ती

हत्ती	From	To	Breaking
CVCCV	0	259	हत्
CVC + CV	259	544	ती

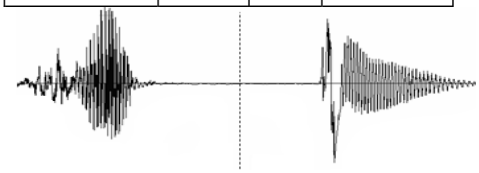


Fig 9 Word हत्ती

In Fig 9 observe that the first syllable हत् has the duration from 0 to 254 and second syllable ती has the duration of 259-

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

544. The plot of sound file also shows that two syllables are distinctly isolated. This is also verified by actually playing and listening the sounds of two syllables. The naturalness of synthetic speech is fairly good. The CV structure breaking rules are sufficient to form enough number of syllables. The syllables thus formed are generating many common as well as uncommon words.

III. SEGMENTATION OF SYLLABLES AND NATURALNESS

This part of work presents methods for automatic speech signal segmentation using neural network. Speech signal segmentation is carried out to form syllables. Concatenative TTS being using speech segments of recorded speech is natural as compare to Formant or Articulatory TTS systems. This TTS stores small segments of speech and join them together to form new word. As manual segmentation is very time consuming and it has certain limitation on naturalness, some neural network models are used to improve naturalness of resulting segments in speech synthesis. About more than 90% accuracy is achieved with neural network models for syllable segmentation which resulted in naturalness improvement of Marathi TTS.

A. Automatic Segmentation of Syllables using Neural Network

Neural Network model for automatic speech segmentation into syllables for 'Devnagari script' is proposed here. Speech synthesizer must be capable of automatically producing speech by storing small segments of speech and splicing and re-splicing them when required.

Neural networks have been applied in speech synthesis for about ten years, and the latest results have been quite hopeful. Automatic segmentation is immensely required because manual segmentation is extremely time consuming. In this work, MAXNET network is explained, which is one layer network that conducts a competition to determine which node has the highest initial input value. Because of one layer, it takes very less time as compared to any other Multi-Layer Neural Network. [4]

B. Feature Extraction

Process of reducing dimensionality of the input is called Feature Extraction. Actual sound file has large number of samples, so input nodes of Neural Network will increase by large number. Therefore Feature Extraction is very important block. Different features can be considered in time and frequency domain. Feature used here is 'energy' which is time domain. It is a simplest and very accurate feature that can be used.

C. MAXNET

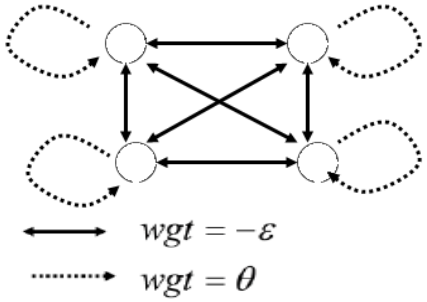


Fig.10 Maxnet Network

Maxnet is simple network to find node with largest initial input value.

Topology: nodes with self-arcs, where all self-arcs have a small positive (excitatory) weight and all other arcs had a small negative (inhibitory) weight.

$\varepsilon \leq 1/(\text{number of nodes})$

$\theta = 1$

Transfer function: fnet = max (net, 0)

$net = \sum_{i=1}^{n}(w_i\, x_i)$ …….. (5)

Basic Algorithm:

Load initial values into the nodes

Repeat: Synchronously update all node values via fnet

Until: all but one node has a value of 0

Winner = the non-zero node

D. SEGMENTATION

Vowel has higher energy as compared to consonants. So syllable has peak in the center and valleys on both the sides. To segment a syllable from word, minima positions should be calculated. Segmentation of 4 syllable word is shown below

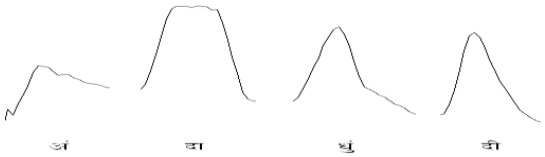


Fig 11 Segmented syllables of अंदाधुंदी

As discussed above, Maxnet is used to find these minima positions. One by one all minimas can be obtained. This gives total segmentation of the word.

E. STORING SEGMENTS

From segmentation, 'from' and 'to' positions of syllable are obtained. These segments need not to be stored separately in audio database. They are stored in textual database.

F. RESULTS OF MAXNET

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

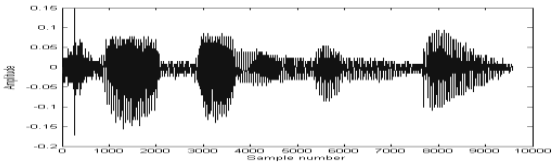


Fig 12 sound file of छेदनबिंदू

TABLE 3
'from' & 'to' positions of छेदनबिंदू

No	Syllable	From	To
1	छे	1	2970
2	द	2971	4400
3	न	4401	6270
4	बि	6271	8910
5	दू	8911	9230

Table 3 shows output of segmentation. 'From' and 'to' positions shown give exact syllable location in that word. Similarly one more result is shown below:

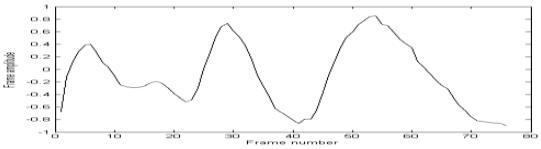


Fig13 Energy plot of बेबंध

TABLE 4
'from' & 'to' positions of बेबंध

No	Syllable	From	To
1	बे	1	2420
2	बं	2421	4510
3	ध	4511	7810

Fig. 13 shows energy plot of word बेबंध and Table 4 shows its syllables.

As unsupervised algorithm, MAXNET, resulted in limited accuracy of resulting minima location/syllable formation hence combination of MAXNET and supervised NN like Back-propagation is implemented.

Back-propagation:

Propagates inputs forward in the usual way, i.e. Propagates the errors backwards by apportioning them to each unit according to the amount of this error the unit is responsible for. MAXNET gives the frame number where minima may be present as output. This output acts as input to Back-propagation. Back-propagation algorithm gives exact sample number/location as minima. This cascaded network of unsupervised and supervised gives more accurate minima position for proper cutting of syllables.

IV. RESULTS OF CASCADED COMBINATION

Example: अडचनी this word contains three syllables, अ + ड + चनी. In Fig. 14 energy plot clearly shows three peaks and three syllables are clearly located.

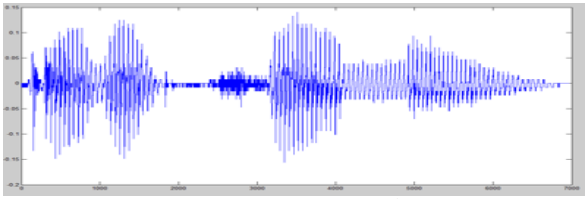


Fig14 Sound File of अडचनी

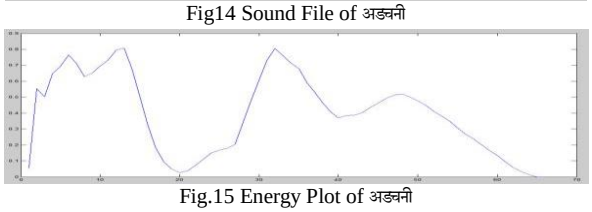


Fig.15 Energy Plot of अडचनी

TABLE 5
Minima Positions of अडचनी

To	From	Breaking
0	2200	अ
2201	4400	ड
4401	6930	चनी

A. GRAPHICAL USER INTERFACE

The naturalness of synthetic speech is fairly good. The CV structure breaking rules are sufficient to form enough number of syllables. The syllables thus formed are generating many common as well as uncommon words. Hence this TTS is working with more naturalness and smoothness by proper syllable cutting or development.

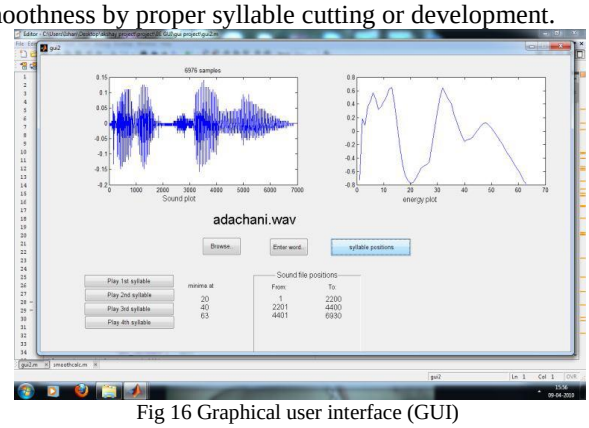


Fig 16 Graphical user interface (GUI)

Neural Network model of both unsupervised and supervised algorithms is implemented accurately for automatic speech segmentation into syllables for MARATHI TTS system. The neural network approaches like Maxnet, K-means outweighs in performance than traditional non neural approaches like slope detection and simulated annealing. With cascaded combination of supervised and unsupervised Neural Network more than 80% accuracy is obtained in speech segmentation. Some more neural network approaches like K-Means along with Maxnet and Backpropagation has given more accurate results. About more than 90% accuracy is achieved with neural network models for syllable segmentation which resulted in naturalness improvement of

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Marathi TTS. Both neural and non-neural approaches are compared in system flowchart for evaluation.

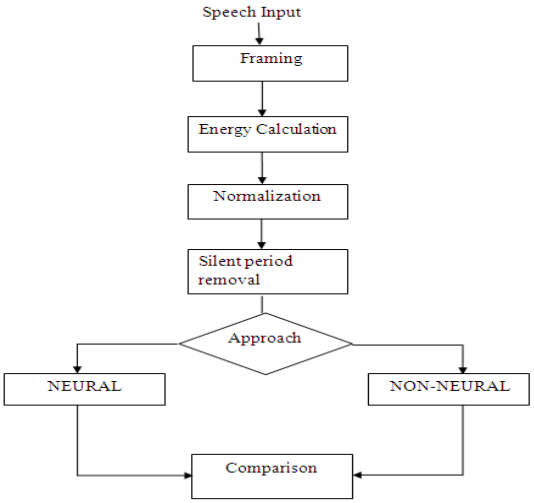


Fig 17 System Flowchart

As shown in the System Flowchart in Fig.17, after energy calculation, normalization and silence removal, both the approaches are compared for the resulting accuracy of segmentation.[6]

B. EXPERIMENTS

For proper breaking of word into syllable there should be one vowel in each syllable. Therefore to separate syllables from word, one requires separating vowels. If one is able to separate two peaks from energy plot, one can separate two vowels.

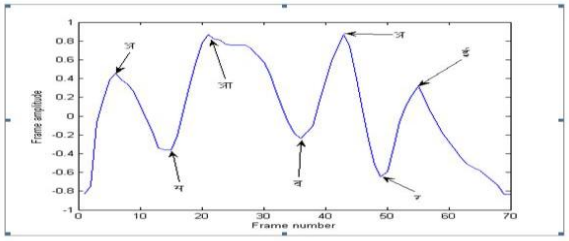


Fig. 18 Energy plot of Marathi word असावरी

Thus, it can be seen from fig 19 that vowels have more energy as compared to consonants. So, three syllables can be segmented from this word.

Here energy plot of word अंदाधुंदी is shown.

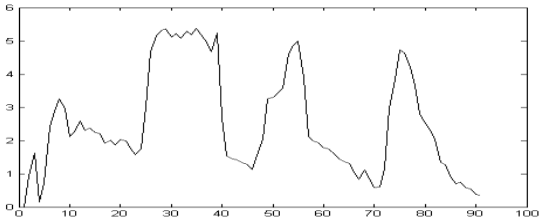


Fig. 19 energy plot of अंदाधुंदी

In this energy plot, many variations are present and also amplitude is not normalized. So amplitude is normalized first from -1 to 1. Then moving average filter is used to reduce variations. See Fig 20 for smooth energy plot of same word.

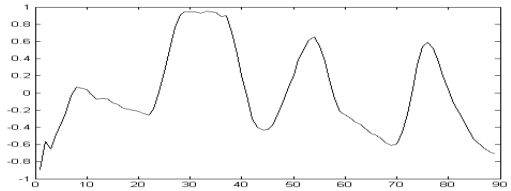


Fig 20 modified energy plot of अंदाधुंदी

Now this modified energy plot is given as input to Neural Network or Non-Neural Approaches.

C. NON NEURAL AND NEURAL ALGORITHMS

1): MAXNET

Maxnet is simple network to find node with largest initial input value. Maxnet gives highest output. So when first 30 frames are applied, output will be the frame number having highest value. Then this frame is discarded and input is again applied to maxnet. After 30 iterations, output will be exact minima. In this way, all minimas can be obtained. Maxnet results are already discussed.

2): K-Means Algorithm

Centroids are to be selected according to the number of syllables in the word.

2. Each centroid contains two parameters a) Amplitude of energy plot b) Frame number.

3. Distance of each point in the energy frame is calculated as

$D=\sqrt{}(x1-x2)\ ^2 + (y1-y2)\ ^2$ …….(6)

4. The energy frame point enters the cluster having minimum distance.

K-Means gives more accurate location of minima or segment point.

3): Simulated Annealing

1. Random Frame numbers are selected.

2. If the energy of next frame number is less as compared with previous then this frame is selected.

3. Temperature variable (T) is used to avoid locking of algorithm in local minimum.

Simulated Annealing gives approximate location of minima or segment cut point.

4): Slope Algorithm

1. The energy plot obtained after normalization and smoothing procedure is used for syllable cutting in non-neural approach.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

2. The objective is to locate the point of inflection on the energy plot where the slope changes from negative to positive.
3. Locating all such points gives segmentation points and hence syllable cutting becomes easier.

5): Results

Maxnet results are already shown above. Results of **Slope Detection** algorithm are shown below for 3 and 4 syllable words.

4-syllable word: अविश्वास

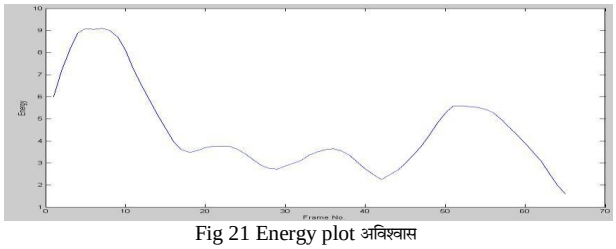


Fig 21 Energy plot अविश्वास

Minima points: Frame number 19, 30, and 43.

3 syllable word: अदिती

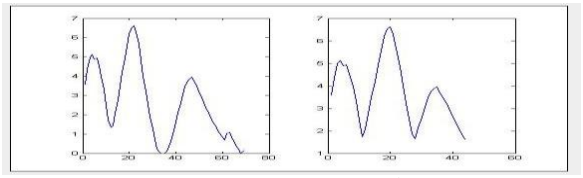


Fig 22 Energy plot of अदिती

It shows the energy graph before and after smoothing. Minima point at frame number 12 and 29.

Results of **K-means**, neural network based algorithm for same words are shown below:

4-syllable word: अविश्वास

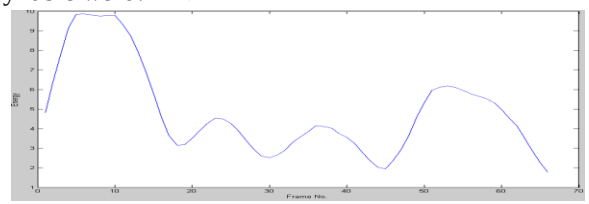


Fig 23 Energy plot of अविश्वास

Minima points: Frame number 18, 30, 45.

3 syllable word: अदिती

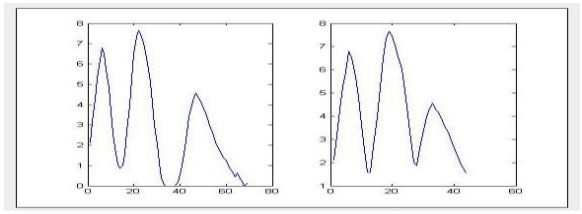


Fig 24 Energy plot of अदिती

Minima point: 13 and 28.

Results of **Simulated Annealing**, one of the nonneural approach are shown below for 2, 3 syllable words.

3-syllable word: अदिती

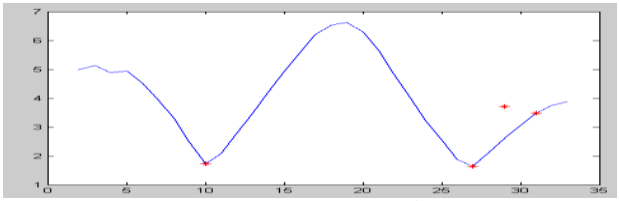


Fig 25 Energy plot of अदिती

Minima points: Frame number: 10, 27

2-syllable word: बातमी

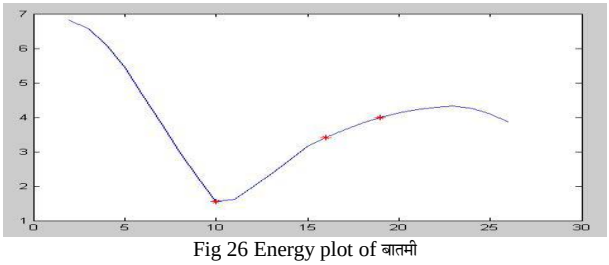


Fig 26 Energy plot of बातमी

Minima Point: Frame Number-10.

Bold numbers in the table indicates that the result is wrong. Comparison of the three algorithms is shown. As can be observed from the table K-means's efficiency is better than the other two algorithms.

TABLE 6
Results for 2-syllables

Word	Simulated Annealing	Slope Algorithm	k-Means
आराम	19	19	19
अब्दुल	9	9	9
उघड	9	8	8
अंकित	16	16	16
चंद्र	28	**23, 28**	28
चळवळ	21	21	21
चिंब	19	19	19
चुंबक	21	21	21
दिशा	**9**	14	14
दुर्गा	17	17	17

TABLE 7
Results for 3-syllables

Word	Simulated Annealing	Slope Algorithm	K-Means
अभिषेक	11,35	11,35	11,35

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

उंबरठा	13,31	13,31	13,31
अमृत	15,25	**15,25,35**	15,25
अमान्य	8,31	8,31	8,31
विदूषी	10,27	**10,27,41**	10,27
चमक्दार	10,27	10,27	10,27
उदासीन	9,29	9,29	9,29
चंदना	**13,19**	19,34	19,34
चुणचुणीत	15,30	15,30	15,30
वस्तुस्थिती	**13,13**	13,29	13,29
चलढकल	29,38	29,38	25,38
चकोरी	14,35	14,35	14,35

TABLE 8
Results for 4-syllables

Word	Simulated Annealing	Slope Algorithm	K-Means
चालुगिरी	24, 44, 61	24, 44, 61	24, 44, 61
चहुकडून	11,29,44	11, 29, 44	11, 29, 44
अविभाज्य	18, 29, 47	**20, 39, 54**	18, 29, 46
अविश्वास	18, 29, 43	18, 29, 43	18, 29, 43
दरोडेखोर	12, 30, 46	12, 30, 46	12, 30, 46
दौऱ्यासाठी	20, 36, 50	20,36, 50	20, 36, 50
देणेघेणे	**38, 50, 56**	18, 38, 56	18, 38, 56

The tabular results for 2, 3 and 4 syllable words shows that Simulated Annealing and Slope Detection algorithm results in minima errors, shown in bold numbers while K-means algorithm is not resulting in error for minima location which shows it's accuracy.

From the results of all four algorithms (neural and non-neural approaches), it is clear that K-means gives more promising results than any other approach. From these results relative functional comparison of these methods can be carried out and hence segmentation accuracy can be decided. The most accurate segmentation method can be used for segmentation of words into syllables.

V. SPECTRAL MISMATCH ESTIMATION IN SYNTHETIC SPEECH

A method based on Power Spectral Density (PSD) to estimate position dependent spectral mismatch is explained. This can be done by plotting power spectral density of 10 millisecond samples of original, properly concatenated (PC) and improperly concatenated (IC) words. These samples are then made noise free to neglect their low amplitude peaks. Analysis of PSD leads to locate formants in the given samples. Formants for original, properly and improperly concatenated words are then plotted. It is observed that formant plots for original and properly concatenated words are very similar for all formants while for improper concatenation extra peaks are observed in all formants. These extra peaks can be considered as estimation for spectral mismatch. The results are validated using Marathi text to speech synthesis. Lot of work is being done around the world in order to achieve a text to speech system (TTS) which sounds as natural as human speech. To design a system which sounds extremely humanoid, the only option is to involve human speech itself. The technique of concatenation is used to form new words from existing syllables in the database. However concatenation approach leads to spectral mismatch at the position of the concatenation of the syllables making synthesized speech more robotic. An estimation of the spectral mismatch is a first step to overcome this problem. The aim of the work is to highlight the spectral mismatch of concatenated syllables due to improper positions using power spectral density. [8]

A. CONCATENATION OF SYLLABLES

The focus of the proposed work is explained with the example.

Original word: देवगड
Concatenated word: Proper concatenation
देवधर + चंदीगड = देवगड
Improper concatenation
रामदेव + गड्गाड्ताना = देवगड

Fig. 27 Example of proper concatenation (PC) and improper concatenation (IC) of word 'देवगड '

The word to be synthesized is देवगड which can be concatenated using syllables देव and गड. The syllable देव can be taken from the word देवधर or from the word रामदेव . Also the syllable गड can be formed from the word चंदीगड or from गड्गाड्ताना. Fig. 27 explains proper and improper concatenation of word देवगड with respect to syllable position.

B. CALCULATION OF PSD

1. After selection of word cut the syllables from words for concatenation.
2. Concatenation can be done manually by using Sound Forge or similar sound editing tools.
3. All the words cut into samples of 10 milliseconds and PSD for each word is plotted.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

4. Peaks are extracted from each and every PSD plot for plotting of formants and slope of each formant is calculated over the range of 50 milliseconds.
5. Normalized values are plotted for each formant for all different combinations and slope of all the formants compared for different combinations.
6. The values of slopes are plotted for more precise combinations.

Fig. 28, 29, 30 below shows PSD of original, PC and IC word 'देवगड '. These PSD plots give 3 formants *f0*, *f1*, and *f2*. The study of these 3 formants for both originally recorded and concatenated word explains the frequency component differences which lead to calculate the spectral mismatch during concatenation.

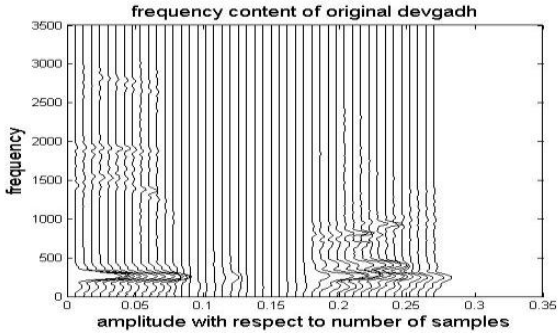


Fig.28 PSD of original 'देवगड '

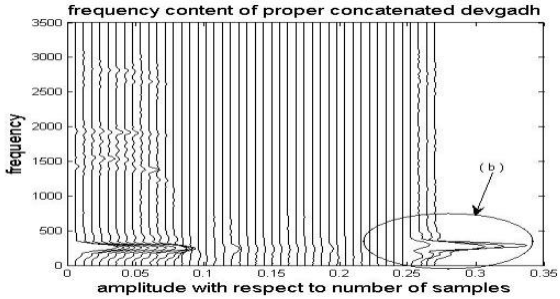


Fig.29 PSD of PC 'देवगड '

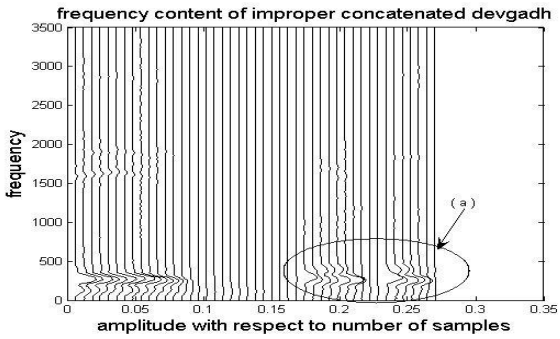


Fig.30 PSD of IC 'देवगड'

In fig. 29 and 30, areas 'b' and 'a' highlights the difference between original, properly concatenated and improperly concatenated word 'देवगड'. The PSD plot in fig. 30 shows sudden decrease in energy at area 'a'. This is because in IC 'देवगड' the second syllable 'गड' has been taken from word 'गड्डाआना'. Of being taken from initial position, the energy of the syllable 'गड' is not as per requirement. This leads to spectral mismatch between original and IC words. The highlighted area in the plot of PC word 'देवगड' as shown in Fig.29 gives higher energy during the syllable 'गड'. Improper selection of the word for concatenation leads to more spectral mismatch. Selection of words and syllable position (context) plays an important role in concatenative speech synthesis. Energy of the word should be same as the original word to achieve a natural voice even after concatenation.

C. COMPARATIVE ANALYSIS OF FORMANTS

Fig. 31, 32 and 33 below shows formant plots of original, PC and IC words 'देवगड '

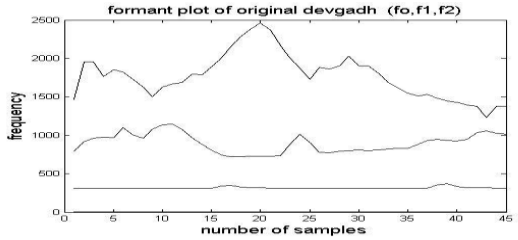


Fig.31 Formant plot of original word 'देवगड'

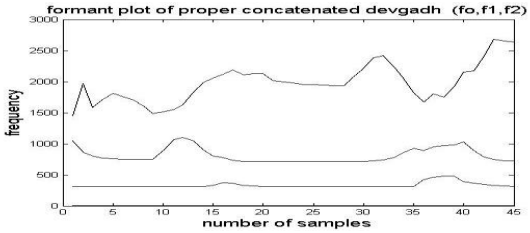


Fig.32 Formant plot of PC word 'देवगड'

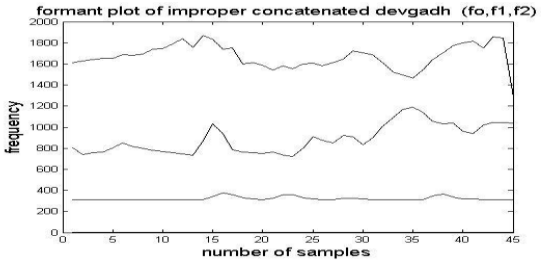


Fig.33 Formant plot of IC word 'देवगड'

The difference in fig. 32 and 33 clearly shows the spectral mismatch, as there is major variation in all the three formants at the concatenation. Study of formants for the same syllable at different positions [initial, final] leads to estimation of spectral mismatch. For this, analysis of each formant is done for original, PC, IC words. Fig. 34 shows a plot of frequency against number of samples for *f0* for original, PC and IC. Fig. 35 and 36 shows similar plots for formants *f1* and *f2* respectively. These figures clearly revels that plot of PC is following the original. However plot of IC

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

shows some extra peaks which clearly indicates spectral mismatch.

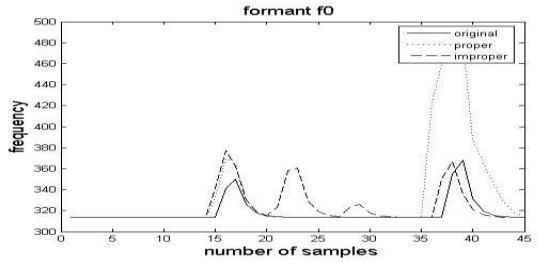


Fig.34 *f0* plot of all combinations

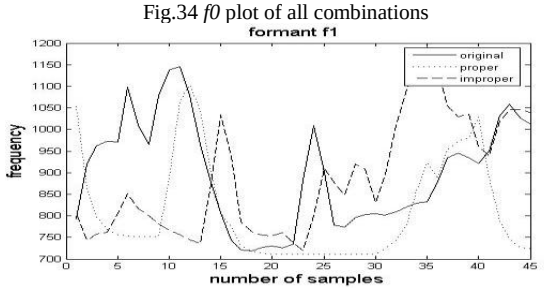


Fig.35 *f1* plot of all combinations

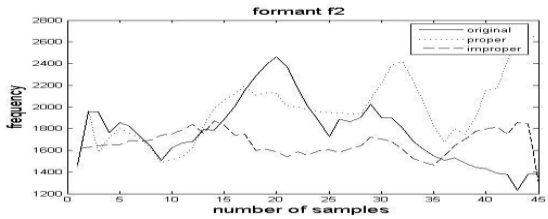


Fig.36 *f2* plot of all combinations

D. SLOPE ANALYSIS OF FORMANTS

The results can be supported by plotting the gradient of formant *f0* over the range of 5 samples for original, PC and IC words. From the plot shown in fig. 37, it can be observed that proper concatenation is much closer to original word than improper concatenation. Similar results can be plotted for *f1* and *f2*.

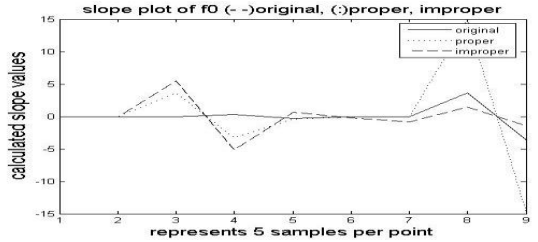


Fig.37 Slope plot for *f0*

The entire discussion supports the fact that improper concatenation show much more spectral mismatch than properly concatenated words. This mismatch can be estimated by Power Spectral Density analysis leading to formant plots. Formant plots of PC are very similar to that of original.

VI. ESTIMATION OF SPECTRAL MISMATCH FOR JOINT COST EVALUATION IN MARATHI TTS

If proper position syllable is used while forming new words from existing syllables, resulting spectral mismatch is less. If position of syllable is not considered during concatenation of speech units, resulting synthesis end up in more concatenation cost. This work presents different techniques like PSD, Wavelet and DTW to find spectral mismatch in concatenated segments. In all these three techniques PSD results are more superior who shows spectral mismatch in graphical form. With direct formant modification one can overcome spectral mismatch and smooth some of the frames which helps to reduce glitch type of sound at concatenation point. Wavelet based audio results shows more naturalness compare to other two methods. This work throws light on how spectral mismatch calculation and reduction increases naturalness of concatenative Marathi TTS. [7]

A. POWER SPECTRAL DENSITY

The power spectral density (PSD) *Sx(w)* for a signal is a measure of its power distribution as a function of frequency. PSD is a very useful tool if one wants to identify oscillatory signals in time series data and want to know their amplitude. Power Spectral Density is used for spectral mismatch calculation.

In proposed work PSD is used to find discontinuities in frequencies and for smoothing.

$$PSD=|x(f)|^2/N \quad \text{.....(7)}$$

Where N is window size.

B. WAVELET TRANSFORM

The analysis of a non-stationary signal using the FT or the STFT does not give satisfactory results. Better results can be obtained using wavelet analysis. One advantage of wavelet analysis is the ability to perform local analysis. Wavelet analysis is able to reveal signal aspects that other analysis techniques miss, such as trends, breakdown points, discontinuities, etc. Here Wavelet analysis is used to find and reduce discontinuities after concatenation. The length of original, proper and improper word is made same. The original, proper and improper word is decomposed up to level 5 with the help of Daubechies Wavelet. The approximate and detail coefficients are extracted from the wavelet transform. For Wavelet transform filter-bank approach is used for which direct MATLAB function is available.

$$\Phi(x) = \sum_{k=-\infty}^{\infty} a_x \ \Phi(S_x - k) \quad \text{.....(8)}$$

The coefficients are given as input to the neural network which is trained with the back propagation algorithm. The output of neural network gives the modified wavelet

coefficients. The original signal is reconstructed from the modified wavelet coefficients.

C. MULTI-RESOLUTION ANALYSIS

For STFT, a fixed time-frequency resolution is used. By using an approach called multi-resolution analysis (MRA) it is possible to analyze a signal at different frequencies with different resolutions. In proposed system multi-resolution analysis is used. The wavelet analysis calculates correlation between the signal under consideration and a wavelet function ϕ(t). The similarity between the signal and the analyzing wavelet function is computed separately for different time intervals, resulting in a two dimensional representation. The analyzing wavelet function ϕ(t) is also referred to as the mother wavelet.

D. MULTI-RESOLUTION ANALYSIS

The continuous wavelet transform is defined as

$$Xxw(\tau, s) = 1/\sqrt{x} \int_{-\infty}^{\infty} x(t)\, \phi\left(\frac{t-\tau}{s}\right) \ldots\ldots\ldots\ldots(9)$$

The transformed signal $Xxw(\tau, s)$ is a function of the translation parameter τ and the scale parameter s. The mother wavelet is denoted as ϕ. The signal energy is normalized at every scale by dividing the wavelet coefficients by 1/√s. This ensures that wavelets have same energy at every scale. The mother wavelet is contracted and dilated by changing scale parameter s. The variation in scale s changes not only the central frequency fc of the wavelet, but also the window length. The translation parameter τ specifies location of the wavelet in time, by changing τ wavelet can be shifted over the signal. The elements in $Xxw(\tau, s)$ are called wavelet coefficients, each wavelet coefficient is associated to a scale (frequency) and a point in the time domain. The inverse continuous wavelet transformation (ICWT) is defined by equation

$$X(t) = 1/c^2\, \phi \iint_{-\infty}^{\infty} x(\tau, s) * \left(\frac{1}{s2}\right) * \phi\left(\frac{t-\tau}{s}\right) \tau\, ds \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

E. DISCRETE WAVELET TRANSFORM

The DWT uses multi-resolution filter banks and special wavelet filters for the analysis and reconstruction of signals.

F. SELECTION OF WAVELET

In comparison to Fourier transform, analyzing function of the wavelet transform can be chosen with more freedom, without the need of using sine-forms. A wavelet function ϕ (t) is a small wave, which must be oscillatory in some way to discriminate between different frequencies. In proposed work Daubechies wavelet is used which is shown in Fig. 38

Fig 38 Daubechies Wavelet

G. DYNAMIC TIME WARPING

The purpose of DTW is to produce a warping function that minimizes the total distance between the respective points of the signals. A concept of an accumulated distance matrix (ADM) is introduced. The ADM contains the respective value in the local distance matrix plus the smallest neighboring accumulated distance. This matrix can be used to develop a mapping path which travels through the cells with the smallest accumulated distances, thereby minimizing the total distance difference between the two signals. This property is used to reduce spectral distance between adjacent frames.

For a distance matrix 'D' difference between each sample value of original and proper word is calculated. Then accumulated distance matrix ADM is calculated with the equation.

ADM(m,n) = LDM(m,n) + min{ADM(m,n-1), ADM(m-1,n-1), ADM(m-2,n-1)} ……………..(11)

ADM= accumulated local distance.
LDM(m,n)= local distance matrix=x(m)-y(n)
The shortest path is found from the 'D' matrix which is used to adjust frames of proper word. After taking inverse FFT, proper word is resized to its original length. Same procedure is repeated for original and improper word.

H. BLOCK DIAGRAM

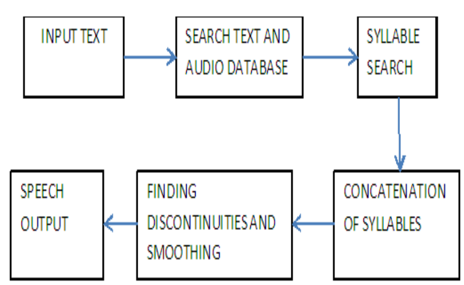


Fig. 39 Block Diagram of System

Important Blocks are Concatenation of Syllables and Finding Discontinuities and Smoothing.

1): Concatenation of Syllable

Concatenations of words are made to generate proper-concatenated word and improper-concatenated words, so as to compare the original word with the concatenated words and find out the difference.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

2): Finding Discontinuities and Smoothing

The project emphasizes the difference between proper and improper concatenated words in text to speech synthesis. Different parameters are used to find discontinuities and spectral smoothing. E.g. Power Spectral Density, Wavelet Transform, Dynamic Time Warping and Back-propagation.

3): Speech Output

Output speech block consist of low pass filter to reduce remaining noise present in synthesis word.

I. SYSTEM FLOWCHART

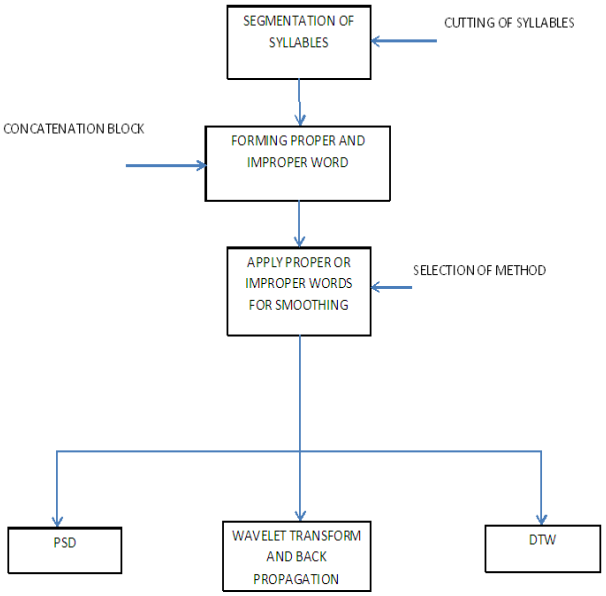


Fig.40 System Flowchart

J. SYSTEM ALGORITHM

1) The input to the text to speech synthesis is Devnagari (Marathi) text. The input word is searched in the database. If word is found it is given to the output file.

2) If word is not found, it is broken into syllables using CV rules. The syllables are then searched into the database and are given to concatenation unit. Proper and improper concatenated word is formed. The discontinuities are found in the concatenated word and they are removed by applying different smoothing techniques. After smoothing the word is given to the output file.

1): Neural Network

Neural networks have been applied in speech synthesis and the results have been quite hopeful. Syllable formation is immensely required because manual formation of syllable is extremely time consuming.

2): Back Propagation

Back-propagation is used to calculate the gradient of error of the network with respect to network's modifiable weights. This gradient is then used to find weights that minimize the error. Here Back-propagation is used along with Wavelet algorithm to reduce spectral discontinuities.

K. RESULTS

1): PSD Results

PSD of each frame of original, proper and improper words is plotted. The cutting frame is shown in red colour. The x-axis is time and y-axis is frequency. Thus it is time-frequency representation i.e. STFT representation.

2): Results for Two Syllable words

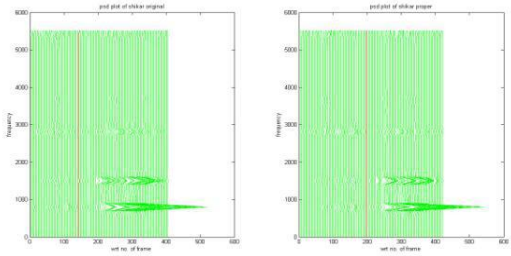


Fig 41 PSD plot of original Shikar and properly concatenated Shikar.

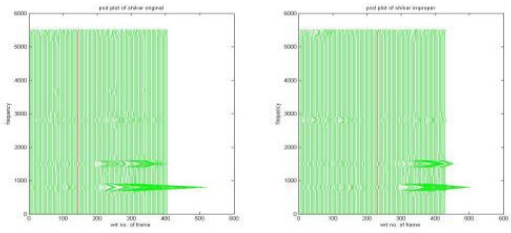


Fig 42 PSD plot of original 'Shikar' and improperly concatenated 'Shikar'.

The part of the plot before red line i.e. cutting frame is syllable 'shi' and the part after the red line is syllable 'kar'. The formants of proper word are similar to the formants of original word.

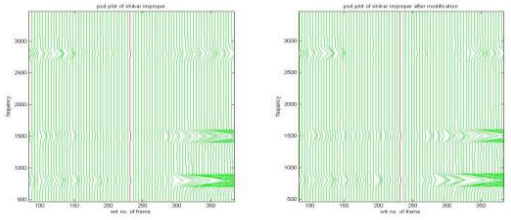


Fig 43 PSD plot of improper 'Shikar' and modified PSD plot of improper 'Shikar'.

In the modified plot shown in fig 43, the formants of improper 'shikar' are modified such that they are made similar to the formats of original 'shikar'.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

3): Numerical Results of PSD

The difference between the formants of original-proper and original-improper words is calculated. The difference is taken for 15 frames before and 15 frames after the concatenation point. The difference is taken for all the three formants. The values in red are the values for concatenation frame.

TABLE 9
Numerical results for word 'Shikar'

First Formant		Second Formant		Third Formant	
Orig-Proper	Orig-Improper	Original-Proper	Original-Improper	Original-Proper	Original-Improper
4.372123	0.045172	1.604309	2.436136	1.251421	2.815694
3.620833	1.506938	0.460083	2.374748	1.989462	4.631546
2.40446	4.511822	3.265818	0.677004	0.489891	1.615047
0.638449	5.801731	0.370185	2.26516	0.922886	1.495002
2.088958	0.81154	2.561768	0.441863	5.143503	0.438593
5.265664	1.476127	0.368604	0.560566	6.485193	0.678317
2.165232	7.76862	0.395312	0.702921	9.625612	0.922336
2.045654	6.626132	4.291382	0.852739	3.669644	3.065797
2.234591	3.023234	3.792063	3.039177	4.369755	0.829948
0.382335	0.231937	2.800457	4.684991	2.901609	1.280343
1.376366	0.218391	0.101408	1.08962	1.307272	0.808039
1.168121	0.671595	1.472119	0.408066	1.01305	0.199581
0.049982	0.774933	1.834192	0.288952	0.431505	0.455103
0.088094	0.116415	0.43646	0.588045	0.174199	0.623503
0.439592	0.614891	0.097197	0.331646	0.508808	0.218166
0.109936	0.132355	0.522631	0.059446	0.106916	0.215197
0.261019	0.294106	0.248495	0.340329	0.291089	0.043552
0.062745	0.92278	0.280008	0.70827	0.544073	0.465108
0.55765	0.023611	0.180804	0.152585	0.117792	0.057613
0.369961	0.02873	0.43877	0.114684	0.131256	0.200727
0.122514	0.437498	0.148938	0.26834	0.008189	0.003048
0.462915	0.050152	0.143341	0.471157	0.120646	0.301773
0.647162	0.687923	9.965161	0.562599	0.330939	0.396689
3.826413	0.287544	7.373661	4.856716	0.155745	0.414947

From the table it can be seen that there is large difference in the formant values of the proper, improper and original words which indicates the spectral mismatch.

4): Results for Three Syllable words

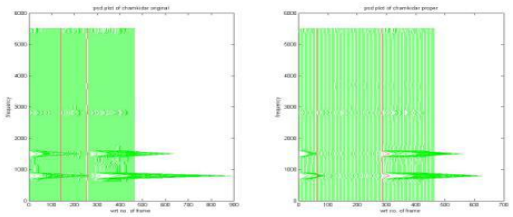


Fig 44 PSD plot of original 'Chamkidar' and properly concatenated 'Chamkidar'.

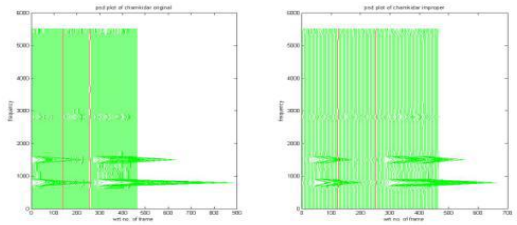


Fig 45 PSD plot of original 'Chamkidar' and improperly concatenated 'Chamkidar'

There are two cutting points in the figure 45 as it is three syllables word. The part of the plot before first cutting point is syllable 'cham', the part after that is syllable 'ki' and the part of the plot after second cutting point is syllable 'dar'.

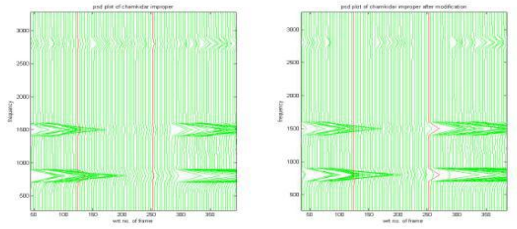


Fig 46 PSD plot of original 'Chamkidar' and improperly concatenated 'Chamkidar' after modification.

5): Numerical Results for PSD

The difference between the formants of original-proper and original-improper words is calculated. The difference is taken for 10 frames before and 10 frames after both the concatenation points.

Numerical results for three syllable word are similar to two syllables except there are two cutting points.

6): Results for Four Syllable Word

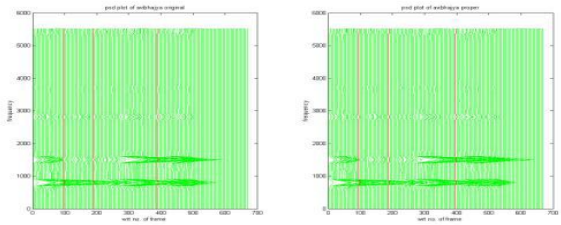
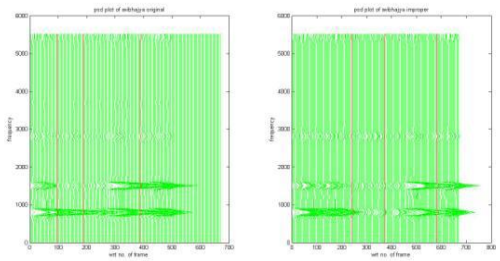


Fig 47 PSD plot of original 'Avibhajya' and properly concatenated 'Avibhajya'



ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Fig 48 PSD plot of original 'Avibhajya' and improperly concatenated 'Avibhajya

In figure 48 there are three cutting points as it is four syllables word. The part of the plot before first cutting point is syllable 'a' , the second part is syllable 'vi', third part is syllable 'bhaj' and the last part is syllable 'ya'.

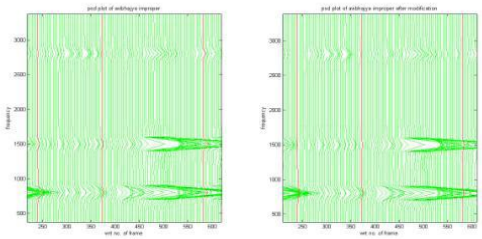


Fig 49 PSD plot of original 'Avibhajya' and improperly concatenated 'Avibhajya' after modification

7): Numerical Results

The difference between the formants of original-proper and original-improper words is calculated. The difference is taken for 5 frames before and 5 frames after three concatenation points.

From all results it can be seen that there is large difference in the formant values of the proper, improper and original words which indicates the spectral mismatch in the proper and improper words. The mismatch is more near the concatenation frame.

L. WAVELET RESULTS

The original, proper and improper word is decomposed up to level 5 using Daubechies wavelet and the wavelet coefficients at each level are plotted for all words. The x-axis shows the coefficient number and the y-axis shows the energy of wavelet coefficients.

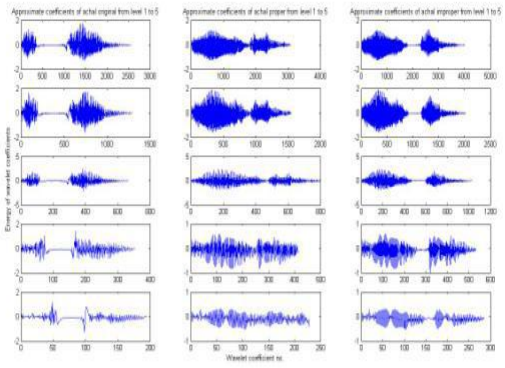


Fig 50 Approximate wavelet coefficients of 'Achal' from level 1 to 5

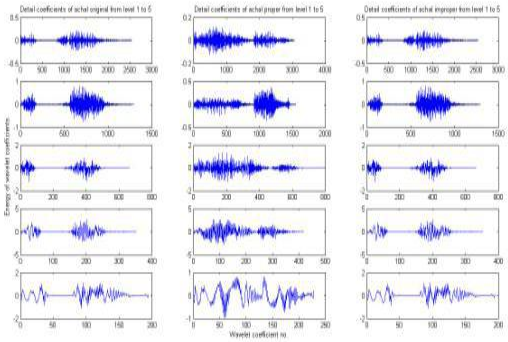


Fig 51 Detail wavelet coefficients of 'Achal' from level 1 to 5

It can be seen from the figure that there is mismatch in the wavelet coefficients of proper and improper word. Leftmost figure is for original word, middle one is for proper and rightmost is for improper concatenated word. Level 1 to 5 indicates energy or amplitude of respective frequency. The mismatch is more at level 5 than the other levels for both approximate and detail coefficients. The corresponding frequency band at level 5 for approximate coefficients is 0 to172 Hz and 172Hz to 344Hz. These coefficients are given as input to the neural network which is trained with the back-propagation algorithm which aims to reduce the mismatch between proper, improper and original wavelet coefficients. [9]

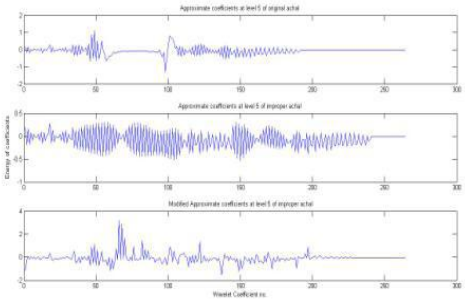


Fig 52 Approximate wavelet coefficients of original, improper and modified improper 'Achal'

It can be seen from the figure 53 that the approximate wavelet coefficients of improper 'Achal' are modified. The neural network has reduced the mismatch between original and improper wavelet coefficients. The reduction in mismatch can be observed with similarity of original and modified improper word after wavelet transform.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

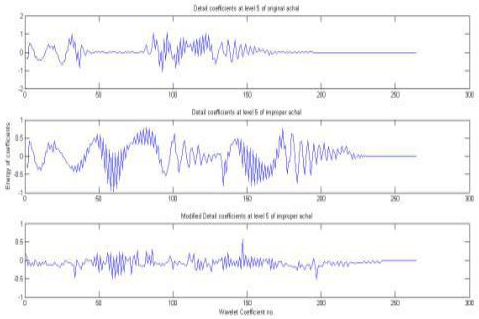


Fig 53 Detail wavelet coefficients of original, improper and modified improper 'Achal'

It can be seen from the figure 53 that the detail wavelet coefficients of improper 'Achal' are modified. The neural network has reduced the mismatch between original and improper wavelet coefficients. The reduction in the mismatch is seen in the % error table.

TABLE 10
Numerical results of Wavelet and Back-propagation for approximate coefficients

Word	Percentage Error Orig-Imp	Percentage Error Orig-M Imp
Geetkar	49.83	28.65
Kamgar	75.42	34.26
Ramdev	44.68	14.66
Marekari	39.24	16.35
Maydesh	63.43	25.26
Savdhan	34.88	16.06
Varkari	51.96	29.57
Upay	24.89	11.84
Vinayak	51.96	13.86

The decrease in the percentage error shows that neural network has reduced the mismatch in the improper word.

M. DTW CORRELATION RESULTS

DTW (Dynamic Time Warping) can be used to improve spectral mismatch in concatenative TTS. The cross-correlation of original and proper and original and improper words is computed before DTW and after applying DTW. The value of correlation increases after DTW. Let Cop and Coi be the cross-correlation between original and proper and original and improper. Then formula used for calculating cross-correlation is

$$\sum_{i=1}^{N} Cop\,(i)/N \quad \ldots\ldots\ldots\ldots\ldots\ldots(12)$$

Following figures and table shows results of correlation

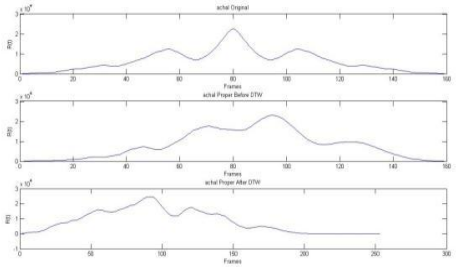


Fig 54 Original and Proper Anchal before and after applying DTW

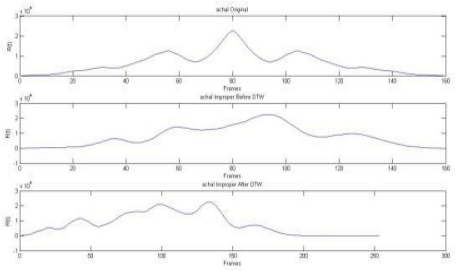


Fig 55 Original and Improper Anchal before and after applying DTW

Above figures shows the correlation results for 'Achal' original, 'Achal' proper before applying DTW and after applying DTW and 'Achal' improper before applying DTW and after applying DTW. Like 'Achal' many other 2, 3 and 4 syllable words are tested. The following table shows the values of the correlation for proper and improper words before and after applying DTW.

TABLE 11
Numerical results of correlation

Word	Cross Correlation of Original and Proper word		Cross Correlation of Original and Improper word	
	Before DTW	After DTW	Before DTW	After DTW
Achal	$2.3056*10^4$	$2.4586*10^4$	$2.2525*10^4$	$2.2587*10^4$
Geetkar	$5.6632*10^3$	$5.269*10^3$	$4.8921*10^3$	$5.8026*10^3$
Varkari	$7.0598*10^3$	$8.6418*10^3$	$9.7344*10^3$	$9.2664*10^3$
Afva	$1.3279*10^4$	$1.6421*10^4$	$1.2571*10^4$	$1.4139*10^4$
Devgad	$1.2314*10^4$	$1.3443*10^4$	$1.081*10^4$	$1.2417*10^4$
Marekari	$7.4595*10^3$	$9.5448*10^3$	$8.0797*10^3$	$8.9536*10^3$
pawder	$1.7639*10^4$	$2.1324*10^4$	$1.3613*10^4$	$1.3823*10^4$
vijay	$9.8615*10^3$	$1.1435*10^3$	$8.5772*10^3$	$8.9659*10^3$
upay	$7.8091*10^3$	$9.1944*10^3$	$6.4715*10^3$	$7.0319*10^3$

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

From the table it can be seen that value of correlation that is similarity of improper and proper word with the original word is increased after applying DTW. Correlation of proper concatenated words is more as compare to improper concatenated words. DTW improves correlation and hence reduces spectral mismatch at concatenation point. All three methods, PSD, Wavelet and DTW help to estimate spectral mismatch (spectral artifacts) and for reduction of such mismatch. [10]

VII. CONCLUSION

From Neural and Non-neural approaches of segmentation, it is clear that k-means gives more promising results than other neural or non-neural algorithms. Accuracy of these algorithms can be judged from tabular results for all types of words (two, three, four syllable words, words with different contexts). From these results relative functional comparison of these methods can be carried out and hence segmentation accuracy can be decided. The most accurate segmentation method is used for segmentation of words into syllables and hence this approach will help to prepare more natural and moderate database TTS system.

Different time and frequency domain methods are tested for spectral mismatch calculation and reduction. In concatenative TTS syllable position plays a very important role. If proper position syllable is used for concatenation of new word (which is not in database) then resulting spectral mismatch is less as compare to improper position syllable used for concatenation. All three methods PSD, Wavelet and DTW show results for both proper and improper syllable position in concatenation. PSD results show numerical difference between original-proper and original-improper words. Spectral distance before and after concatenation point can be clearly seen in PSD graphs. Wavelet with back-propagation algorithm improves mismatch of improper concatenated word. Numerical results of DTW shows increase in correlation value after applying DTW. Correlation of proper concatenated words is more as compare to improper concatenated words. DTW improves correlation and hence reduces spectral mismatch at concatenation point. In future work, wavelet based spectral mismatch reduction can be extended further to improve audio results. PSD results have limitation of graphical form and resulting audio performance is limited. DTW being time domain parameter, accuracy is not up to the mark, in future work it can be improved with some frequency domain parameter. Thus with accurate segmentation and spectral mismatch reduction after concatenation, naturalness of proposed Marathi TTS can be increased.

REFERENCES

[1] "Objective distance measure for spectral discontinuities in concatenative speech synthesis."—J.Vepa, S. King and P. Taylor, in proc. ICSLP, Denver, co, 2002.

[2] "The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation" – T. Nagarajan, V. Kamakshi Prasad and Hema A. Murthy, Sixth Biennial conference of signal processing and communications, July 2001.

[3] "A comparision of spectral smoothing methods for segment concatenation based speech synthesis", -David T. Chappell, John H. L. Hansen.

[4] "Context-Adaptive Smoothing for concatenative speech synthesis", - Ki-Seung Lee and Sang-Ryong Kim, IEEE signal processing letters, vol.9, No. 12, December 2002.

[5] "Refining segmental boundaries for TTS Database using fine contextual dependent boundary models", - Lijuan Wang, Yong Zhao, Min Chu, Jianlai Zhou and
Zhigang Cao.

[6] " Subjective evaluation of joint cost and smoothing methods for unit selection speech synthesis", - Jithendra Vepa and Simon King, IEEE transactions on Audio, Speech, and Language Processing, Vol. 14, No.5, September 2006.

[7] "New Objective Distance measures for Spectral Discontinuities in Concatenative speech synthesis.", - Jithendra Vepa, Simon King and Paul Taylor, IEEE 0- 7803-7395-2/2002.

[8] "Concatenative Speech Synthesis for European Portuguese", - Pedro M. Carvalho, Luis C. Oliveira, Isabel M. Trancoso, M. Ceu Viana, INESC/IST. 2006

[9] "Sub-band based group delay segmentation of spontaneous speech into syllable like units", -T. Nagarajan, H.A. Murthy, I.I.T. Madras. 2008

[10] "Concatenation cost calculation and optimization for unit selection in TTS", -christophe Blouin, Oliver Rosec, Paul c. Bagshaw and Christophe d'Alessandro, IEEE-0-7803-7395-2/2002.