Relative Functional Comparison of Neural and Non-Neural Approaches for Syllable Segmentation in Devnagari TTS System.

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

[1] Research Scholar, Bharati Vidyapeeth College of Engineering, Pune, 411043, India.

[2]Dept of Electronics, Bharati Vidyapeeth College of Engineering, Maharashtra, Pune, 411030, India.

Abstract

This paper presents methods for automatic speech signal segmentation using neural network. Speech signal segmentation is carried out to form syllables. Syllable is a common unit for concatenative TTS systems. 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. This helps to generate more number of words based on very small database. 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. The proposed work explains how neural network approaches like Maxnet, K-means outweighs in performance than traditional non neural approaches like slope detection and simulated annealing. About more than 90% accuracy is achieved with neural network models for syllable segmentation which resulted in naturalness improvement of Marathi TTS.

Keywords: Neural Network Approach, Non-Neural Approach, Text to Speech (TTS) System, Speech Segmentation.

1. Introduction

A text-to-speech (TTS) system converts some particular language text into speech. Speech synthesis is the artificial production of human speech. Synthesized speech can be created by concatenating pieces of pre-recorded speech that are stored in a database. The most important qualities of a speech synthesis system are naturalness and intelligibility. The key objective of proposed work is to design a system that develops syllables automatically from different words. Proper syllabification is required so that each syllable can be used in formation of new word with improved naturalness of synthetic speech. Use of neural networks has been introduced for syllable cutting. The importance of neural network in syllabification is:

1. Its ability to generalize and capture the functional relationship between input and output pattern pairs.
2. Its ability to predict, after an appropriate learning phase, even those patters not presented before.
3. Its ability to tolerate certain amount of fault in input.

Speech synthesizer must be capable of automatically producing speech by storing small segments of speech and splicing and re-splicing them when required. The syllable serves as an important interface between the lower level and the higher-level representational tiers of language. It generates more number of words based on very small database. Different syllables can form new words hence original database is not large. 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 based Approach provides automatic speech segmentation. Algorithms like Maxnet and K-means are used in proposed work for carrying out proper segmentation of speech. To compare the accuracy of neural network with that of some non-neural approaches, two approaches like Slope Detection and Simulated Annealing are discussed. Maxnet and K-Means are providing more than 90% accuracy in syllable segmentation. While the accuracy of non-neural approaches like Slope Detection and Simulated Annealing is limited to less than 70%. The main objective is to develop small segments (i.e. syllables) from different words. Proper cutting of words (i.e. segmentation of syllables) is required so that each syllable can be used in the formation of new word, which increases naturalness of synthetic speech.

1.1 Syllable

A ***syllable*** is a phonetic-phonological basic unit of a word. It consists of syllable start, syllable core and syllable end. Syllables are not influenced by neighbored sound elements. The segmentation of syllables is relatively easy.

e. g.: nmUr = nm + Ur

Here nm and Ur are two syllables. Basically, syllable is combination of consonants and vowels.

1.2 Neural Network

Automatic segmentation is required because manual segmentation is extremely time consuming. Also manual segmentation will have some restrictions because of human limitations. In this paper, MAXNET and k-means algorithms are explained. MAXNET is one layer neural 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 Back- propagation or any other Multi-Layer Neural Network. For K-means the centroid was selected so that minima lies in the 2nd cluster or third cluster depending on number of syllables in the given word. Then using MAXNET minima can be located.

1.3 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 in this paper is 'energy' which is in time domain form.

1.4 Energy Calculation

Normally the energy of the signal is defined as

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

where m varies from zero to plus infinity.

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

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

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 the 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 sample-to-sample variation in x(m). A simple way to lessen the effect of this problem is to use average magnitude function for calculating energy. The formula for this is given as

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

where m varies from n-N+1 to n. This function is called average magnitude but here it is called as energy only. Every frame is formed of 10msec of sound file. This is because of limitation of Human Vocal system. Higher frequency than this can't be produced by humans. The recording frequency used was 11025 Hz. So 10msec frame will have 110 samples of sound file.

2. Literature Review

Speech is the most natural and efficient form of communication and provides a vehicle with which to transmit our thoughts and ability to impart large amount of information in a compact form. Naturalness in human speech is dependent on a number of factors and the extent to which a text to speech synthesis system can account for these factors in its model will be a measure of its success in the marketplace.

Various speech synthesis methods are available to improve the speech quality and naturalness for TTS systems. A few of them are reviewed here.

2.1 Context Based Speech Synthesis

Speech in Indian language is based on basic sound units which are inherently syllable unit made from C, V, CCV, VC and CVC combinations, where C is a consonant and V is a vowel. The syllable like unit is defined in terms of the short-term-energy (STE) function by M. Nageshwara Rao et.al. [1]. To create syllable like unit repository, a group delay based segmentation algorithm can be used.

Segmentation of acoustic signal into syllabic units is important stage in speech synthesis. Although STE (Short Term Energy) function contains useful information about syllable segment boundaries, it has to be processed before segment boundaries can be extracted. T. Nagarajan et.al. [2] presents a sub-band based group delay approach to segment spontaneous speech into syllable-like units.

Lijuan Wang et.al. [3] presents a post-refining method with fine contextual-dependent GMMs for auto-segmentation task. A GMM trained with a super feature vector extracted from multiple evenly spaced frames near the boundary is suggested to describe the waveform evolution across a boundary.

S. P. Kishore et.al. [4] describes a data-driven synthesis for Indian languages using syllables as basic units for concatenation. Unit selection algorithm of this method exploits the advantages of a prosodic matching function, which is capable of implicitly selecting a larger sequence such as words, phrases and even sentences.

The mark-up of stored waveforms for segmentation into syllables must be precise; it can be done manually or automatically. Eric Lewis et.al. [5] describes how most of the segmentation can be done automatically, leaving only those waveforms which would be prone to error to be segmented manually.

Goh Kawai et.al. [6] suggests a method for detecting syllabic nuclei from English utterances on a frame-by-frame basis using band pass filtered acoustic energy measurements.

2.2 Evaluation Of Speech Synthesis With Spectral Smoothing Methods

There are many scenarios in both speech synthesis and coding in which adjacent time-frames of speech are spectrally discontinuous. David T. Chappell et.al. [7] addresses the topic of improving concatenative speech synthesis with a limited database by proposing methods to smooth, adjust, or interpolate the spectral transitions between speech segments. Techniques examined include optimal coupling, waveform interpolation, linear predictive parameter interpolation, and psychoacoustic closure.

Previous studies on joint cost have focused predominantly on static spectral measures extracted from the unit boundary. Spectral dynamic behavior can be investigated as a source of discontinuity in concatenated speech. This method is suggested by Barry Kirkpatrick et. al. [8].

A context adaptive smoothing method is proposed by Ki-seung Lee et. al. [9], where amount of smoothing is determined according to context information. Discontinuities at unit boundaries are predicted by a regression tree and smoothing factors are computed by using predicted discontinuities and real discontinuities at unit boundaries.

Usually, some form of local parameter smoothing is also needed to disguise the remaining discontinuities. Jithendra Vepa et. al. [10] presents a subjective evaluation of three join cost functions and three smoothing methods.

The objective measure of discontinuity used when selecting units is known as the join cost. Jithendra Vepa et. al. [11] described a perceptual experiment conducted to measure the correlation between subjective human perception and various objective spectrally-based measures proposed in the literature.

2.3 Concatenative Speech Synthesis And Segmentation

A concatenative speech synthesis system consists of three components. The first component, the text analysis frontend, takes text as input and outputs a sequence of feature vectors that characterize the acoustic signal to synthesize. The second component, unit selection determines in a set of recorded acoustic units corresponding to phones or half-phones the sequence of units that is closest to the sequence of feature vectors predicted by text analysis front end. The final component produces an acoustic signal from the unit sequence chosen by unit selection using simple concatenation or other methods.

Duration is one of the prosodic feature of speech, the other two being stress and intonation. S. R. Rajesh Kumar et. al. [12] demonstrates the significance of durational knowledge in speech synthesis for the Indian language, Hindi.

Speech synthesis systems based on concatenation of segments derived from natural speech are very intelligible and achieve a high overall quality. Matthias Eichner et. al. [13] suggests a new approach for duration control in speech synthesis that uses the probability of a word in its context to control the local speaking rate within the utterance.

S. P. Kawachale et. al. [14] describes syllable based speech synthesizer. 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. Dictionary based synthesis has the drawback of 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.

S. P. Kawachale et. al. [15] shows that syllabic based speech synthesizer uses energy calculation technique. While storing small segments syllable cutting becomes essential part in concatenative type speech synthesis. Different syllables can form new words hence original database is not large. For properly cutting of syllables location of vowels plays an important role. 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.

K. Szklanny et. al. [16] reports progress in the process of improving automatic segmentation. Process of automatic segmentation often causes errors to be corrected. Script for finding outliers was constructed and only necessary manual correction was proposed. Then a praat script was realized which allowed to detect and remove most important errors in automatic segmentation.

With the growing popularity of corpus-based methods for concatenative speech synthesis, a large amount of interest has been placed on borrowing techniques from ASR community. Ivan Bulyko et. al. [17] explores the applications of Buried Markov Models (BMM) to speech synthesis. The paper shows that BMMs are more efficient than HMMs as a synthesis model, and focus on using BMM dependencies for computing splicing costs.

Several new techniques that improve text-to-speech synthesis by using neural network have been developed. Rachid Hamdi et. Al. [18] explains speech synthesis using optimized Neural Networks with Genetic Algorithms.

3. Proposed Methodology

The main objective of the research can be achieved by exploring sub-objectives or methodologies.

3.1 Design of complete Framework for Speech Synthesis-by-concatenation

A complete framework includes a Speech Synthesis Engine (developed in C / Matlab platform) with a GUI front end. (developed in Visual Basic / Matlab). A GUI keyboard allows inputting the Marathi Text. Input text is then passed to Speech Synthesis engine search block.

Text analysis block aims to carry out text encoding and search through text strings. For syllabification, all the CV rules need to be studied. Based on CV structure rules and using STE (short term energy) algorithm, syllable formation need to be carried out. More naturalness can be achieved if syllable search is based on IMF (Initial, Middle and Final) position. Hence context based database formation and segmentation (syllable cutting from words) has to be implemented. After segmentation to check the performance of resulting speech output, calculation of spectral mismatch has to be done.

3.2 Study of linguistic properties of Marathi to select 5000 words for database

To improve the overall performance of current Marathi TTS, rich database (vocabulary) needs to be prepared. For designing rich database, linguistic properties of Marathi language should be studied. Also CV (consonant-vowel) rules need to be checked for most frequent contexts. Based on this study, a database of around 4000 to 5000 words has to be prepared.

3.3 Preparation of database of 5000 words and syllables

While preparing database of sufficient vocabulary, words or speech corpora needs to be selected based on language study. As current Marathi TTS is based on concatenative speech synthesis, most frequently used words need to be recorded. The recording has to be carried out with noise resistant microphone and with standard multimedia PC. Also the recording should be done in a sound-proof room.

3.4 Context based speech synthesis:

While forming new word, if position of syllable is considered (context), then few rules can be applied to concatenation program and more natural output can be achieved. The syllable positions (initial, middle and final) give different values for intonation, loudness and also their prosody is different. If syllable at position final is used at initial position, then it will result in longer delay which is not required for initial position. If other positions of syllables are changed in a similar fashion then we may get un-natural audio output.

3.5 Improving spectral smoothing

The naturalness of the synthesized speech is strongly influenced by the match between elements concatenated together. Spectral smoothing is an objective measure of naturalness and performance of speech synthesis system. To improve the quality of speech if spectral mismatch is calculated and based on that if concatenation is carried out the audio output may sound more natural.

4. Simulation

4.1 Block Diagram

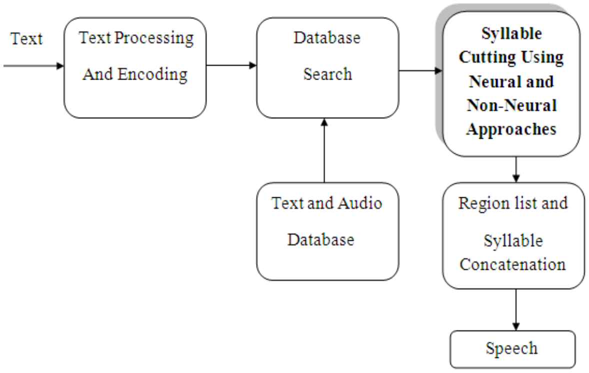

Text
Text Processing And Encoding
Database Search
Syllable Cutting Using Neural and Non-Neural Approaches
Text and Audio Database
Region list and Syllable Concatenation
Speech

Fig.1: Block Diagram

4.2 Explanation of Block diagram

1. The input to the TTS system (Text to Speech Synthesizer) is Devnagari (Marathi) text.
2. In text processing, all spaces, tabs, commas, full stops etc. will be removed and words will be isolated from each other.
3. Encoding stage encodes input text into ASCII code string. This block also breaks word into syllables using CV (consonant-Vowel) rules.
4. Encoded word will be searched in database. Two databases need to be maintained viz. Audio database and Textual database.
5. If word is not found in database then it will be cut(break) into syllables.
6. Syllable cutting using neural and non-neural approaches is performed.
7. After syllable cut the concatenation stage concatenates syllables and words and produces full sentence in speech format.

4.3 System Flowchart

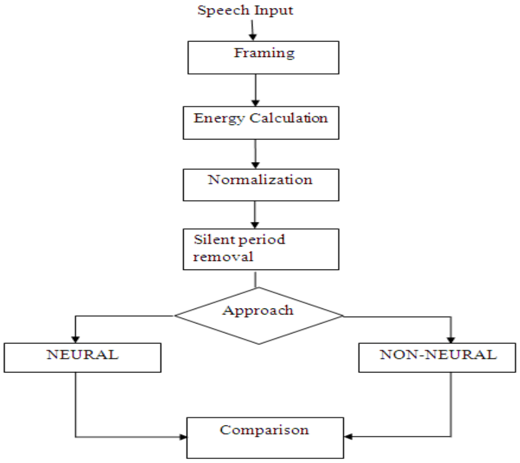


Fig 2: System flowchart

As shown in the System Flowchart in Fig. 2, after energy calculation, normalization and silence removal, both the approaches are compared for the resulting accuracy of segmentation.

4.4 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.

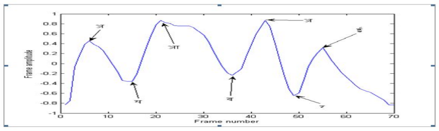


Fig. 3: Energy plot of Marathi word Agmdar

Thus, it can be seen from fig 3.that vowels have more energy as compared to consonants. So, three syllables can be segmented from this word.

4.5 Algorithm for energy calculation

1. Read sound file (.wav format).
2. Take absolute value of every sample.
3. Calculate total number of frames having 110 samples per frame.
4. Sum of absolute values of 110 samples is taken for particular frame. (as formula shown above).
5. Such a calculation for every frame gives energy plot.
6. Energy plot is plotted as frame no. Vs. frame amplitude.

Here energy plot of word AXmYXr is shown.

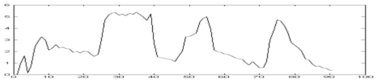


Fig 4: energy plot of AXmYXr

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 5 for smooth energy plot of same word.

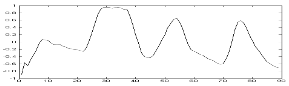
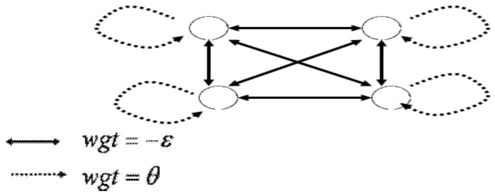


Fig 5: modified energy plot of AXmYXr

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

5. Non-neural and Neural Algorithms

5.1 MAXNET



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.

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

Transfer function: f_{net} =max(net, 0)

$$net = \sum_{i=1}^{n}(Wi * Xi) \tag{4}$$

Basic Algorithm: Load initial values into the nodes

```
   Repeat:
        Synchronously update all node values via
f_net
   Until: all but one node has a value of 0

   Winner = the non-zero node.
```

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. From Fig 5, we can get 4 minima positions because it has 4 syllables. Accordingly we can get segmentation as shown in Fig 6.

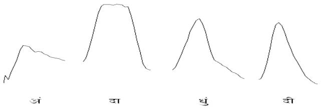


Fig 6: Segmented syllables of AX|Y|Xr

As discussed above, Maxnet is used to find these minima positions. 30 frames are given as input to maxnet at a time. From results it is seen that mostly 30 frames contain a syllable. Once minima is obtained, then next 30 frames starting from calculated minima are given to maxnet.

Maxnet gives highest input. So when first 30 frames are appplied, output will be the frame number having highest value. Then this frame is discarted and input is again applied to maxnet. In that case second maxima will be the output. After 30 iterartions, output will be exact minima. In this way, all minimas can be obtained. This gives the total segmentation of the given input word.

5.2 K-Means Algorithm

1. 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

$$D=\sqrt{(x1-x2)^2 + (y1-y2)^2} \quad (5)$$

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

5.3 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.

5.4 Slope Algorithm

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

6. Results

Results of **Maxnet**, one layer neural network are shown below for few Marathi words.

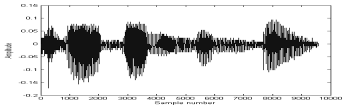


Fig 7: sound file of N|XZþ~X|

No.	Syllable	From	To
1	N|	1	2970
2	X	2971	4400
3	Z	4401	6270
4	þ~	6271	8910
5	X|	8911	9230

Table 1: ‘from’ & ‘to’ positions of N|XZþ~X|

Table 1 shows output of segmentation. ‘From’ and ‘to’ positions shown give exact syllable location in the given word. Similarly some other results are shown below:

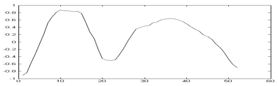


Fig 8: Energy plot of A|^| i

No	Syllable	From	To
1	A|	1	2420
2	^| i	2421	5720

Table 2: ‘from’ & ‘to’ positions of A|^| i

Fig. 8 shows energy plot of A|^| i and Table 2 shows its ‘from’ and ‘to’ positions in sound file.

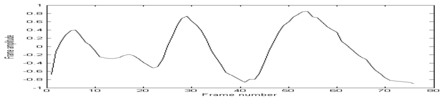


Fig 9: Energy plot of ~|~|Y

No.	Syllable	From	To
1	~|	1	2420
2	~|	2421	4510
3	Y	4511	7810

Table 3: ‘from’ & ‘to’ positions of ~|~|Y

Fig. 9 shows energy plot of word ~|~|Y and Table 3 shows its syllables. Fig. 10 shows energy plot of

5 syllable Marathi word and Table 4 gives its syllable positions.

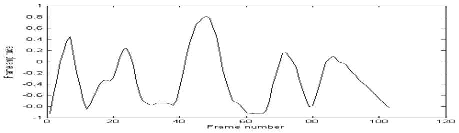


Fig. 10: Energy plot of word M_H XmanZ

No.	syllable	from	to
1	M	1	1320
2	_H	1321	4180
3	Xma	4181	6820
4	n	6821	8690
5	Z	8691	11330

Table 4 ‘from’ and ‘to’ positions of M_H XmanZ

Results of **Slope Detection** algorithm are shown below for 2, 3 and 4 syllable words:

4-syllable word: Apdídmg

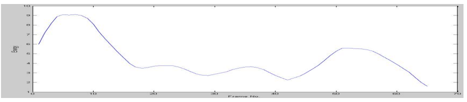


Fig 11: Energy plot Apdídmg

Minima points: Frame number 19, 30, 43.

3 syllable word: ApXVr

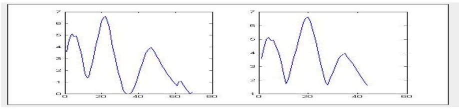


Fig 12: Energy plot of ApXVr

It shows the energy graph before and after smoothing. Moving average filter is used for smoothing of energy plot. Minima point at frame number 12 and 29.

2-Syllable word: MmJb

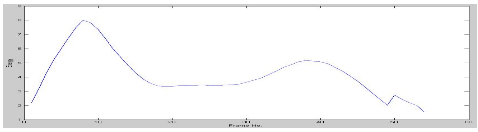


Fig 13: Energy plot of MmJb

Minima points: frame number 20, 27.

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

4-syllable word: Apdídmg

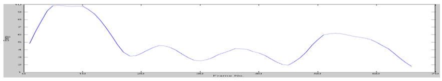


Fig 14: Energy plot of Apdídmg

Minima points: Frame number 18, 30, 45.
Amplitude: 3.1578, 2.5109, 1.9516.

3 syllable word: ApXVr

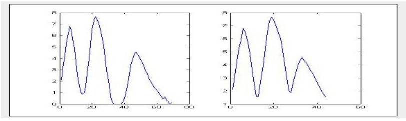


Fig 15: Energy plot of ApXVr

Minima point: 13 and 28.

2-Syllable word: MmJb

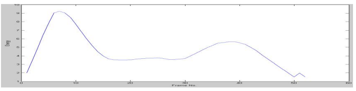


Fig 16: Energy plot of MmJb

Minima points: frame number 19.
Amplitude: 3.5031.

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

3-syllable word: ApXVr

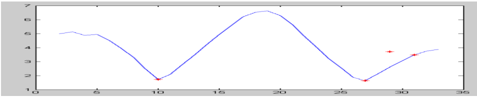


Fig 17: Energy plot of ApXVr

Minima points: Frame number: 10, 27

2-syllable word: ~mV_r

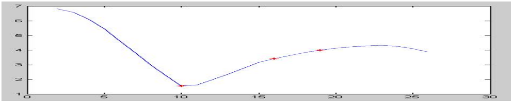


Fig 9.Energy plot of ~mV_r

Minima Point: Frame Number-10.

Results for 2-syllables

Word	Simulated Annealing	Slope Algorithm	K-Means
Amam_	19	19	19
AH a	31	31	31
AmË_m	19	19	19
emYH	29	29	29
OdmZ	16	16	16
ApH V	16	16	16
MÖ	28	**23, 28**	28
M i d i	21	21	21
AmOmar	13	13	13
XOmi	21	21	21
Xmam	17	17	17
pXem	**9**	14	14

Results for 3-syllables

Word	Simulated Annealing	Slope Algorithm	K-Means
Ap^fH	11,35	11,35	11,35
A[YHmma	14,30	14,30	14,30
AmMm`	20,48	20,48	20,48
C~aRm	13,31	13,31	13,31
A_V	15,25	**15, 25, 35**	15,25
A_mY`	8,31	8,31	8,31
pdXfr	10,27	**10, 27, 41**	10,27
MXZm	**13, 19**	19,34	19,34
X[jU_Ir	11,45	11,45	11,45
MUMUrV	15,30	15,30	15,30
dñVpñWVr	**13, 13**	13,29	13,29

Results for 4-syllables

Word	Simulated Annealing	Slope Algorithm	K-Means
MmbpJar	24,44,61	24 44 61	24 44 61
MhH SZ	11,29,44	11 29 44	11 29 44
~JmbgmRr	21,44,61	21 44 61	21 44 61
Apd^mA`	18,29,47	18 29 47	18 29 47
Apdídmg	18,29,43	18 29 43	18 29 43
XZpXZr	17,40,51	17 40 51	17 40 51
XamS I ma	12 30 46	12 30 46	12 30 46
Xmè`mgmRr	20 36 50	20 36 50	20 36 50
XUKU	**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.

7. Conclusion

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. Maxnet results are not very accurate; it just provides correct frame number from energy graph where minima (segment point) lies. The accuracy of these algorithms can be judged with tabular results shown above for three approaches, Slope Detection, K-Means and Simulated

Annealing. 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 and hence this approach will help to prepare more natural and moderate database TTS system.

8. 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.

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

[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.

[11]“A Data driven synthesis approach for Indian languages using syllable as basic unit”, - S. P. Kishore, Rohit Kumar, Rajeev Sangal.

[12]“Text to Speech synthesis using syllable like units.”- M. Nageshwara Rao, Samuel Thomas, T. Nagarajan and Hema A. Murthy, I. I. T. Chennai.

[13] “Improved Duration Control for speech synthesis using a multigram language model”, -Matthia Eichner, Matthias Wolff and Rudiger Hoffman, IEEE-0-7803-7402-9/2002.

[14] “An Optimized Soft Cutting Approach To Derive Syllables From Words In Text To Speech Synthesizer”, -S. P. Kawachale, J. S. Chitode, IASTED, SIP 2006, Vol. No 534, No. 195.

[15]“Identification of Vowels in Devnagari Script Using Energy Calculation”, S. P. Kawachale, J. s. Chitode, Emerging trends in engineering and technology, PCCOE, Goa.

[16] “Automatic segmentation quality improvement for realization of unit selection speech synthesis”, - K. Szklanny, M. Wojtowski, Multimedia Department, Polish-Japanese Institute of Information Technology, HIS, Krakow, Poland, May 25-27, 2008.

[17]“Automatic detection of syllabic nuclei using acoustic measures”, -Goh Kawai and Jan Van Santen, IEEE, 0-7803-7395-2/2002.

[18] “ Arabic speech synthesis using optimized neural networks with Genetic Algorithms” Rachid Hamdi and Mouldi Bedda. Asian Journal of Information Technology 5(7): 686-690, 2006 © Medwell Online 2006.

First Author Prof Mrs Smita P. Kawachale, Sr Lecturer, Maharashtra Institute of Technology, Pune, India, Research Student, Bharati Vidyapeeth COE, Pune, India, ISTE member, 10 national and international conference papers, 4 IEEE sponsored paper presentations.

Second Author Prof Dr Janardan S. Chitode, Honarray professor, Bharati Vidyapeeth College of Engineering, Pune, India, Best author for Technical publication, Research guide, many papers in research journals and conferences.