Estimation of Spectral Mismatch for Joint Cost Evaluation in Marathi TTS

Smita P. Kawachale
Research Scholar
Dept of Electronics, Bharati Vidyapeeth, COE
Maharashtra, Pune, 411030, India

Janardan S. Chitode, PhD.
Honorary Professor
Dept of Electronics, Bharati Vidyapeeth, COE
Maharashtra, Pune, 411030, India

ABSTRACT
Among different methods of speech synthesis, Concatenative Speech Synthesis is widely used due to its naturalness and less signal processing requirement. But concatenative TTS 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 paper 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 we 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. 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. This work throws light on how spectral mismatch calculation and reduction increases naturalness of concatenative Marathi TTS.

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

1. INTRODUCTION
A text-to-speech (TTS) system converts some particular language text into speech. Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer. The ideal speech synthesizer should be natural as well as intelligible. The two primary techniques for generating synthetic speech are formant synthesis and concatenative synthesis. The synthesis method used in proposed work is concatenative synthesis. The unit used for concatenation is syllable. The segmentation of syllables is relatively easy.

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

here (vaar) वार and (kari) क री are two syllables. Basically, syllable is combination of consonants and vowels. One of the major reasons for considering syllable as a basic unit for segmentation is its better durational and representational stability as compared to other units. In proposed system a small database is used. If a word (vaarkari) वारकरी is not found in the database then it is formed by joining the syllables (vaar) वार and (kari) करी from the database.

In case of concatenative synthesis, when a new word is formed, the syllables required for making the word are taken from other words stored in the database. These words are segmented and the required syllables are extracted. The extracted syllable should be present in the same position (initial, middle or final) as required by the new word. This will make synthesized word sound more natural. During concatenation of syllables formants and other spectral characteristics are disturbed which results in spectral mismatch as shown in [1] [3]. The key objective of proposed work is to increase the naturalness of the text to speech system by removing discontinuities occurring during concatenation by implementing different spectral smoothing techniques.

1.1 Power Spectral Density
The power spectral density (PSD) $S_x(w)$ 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.

1.2 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.

1.3 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.

1.4 Continuous Wavelet Transform
The continuous wavelet transform is defined as

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

---------- (1)

The transformed signal XWT (τ , 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 $1/\sqrt{s}$. 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. Therefore the scale s is used instead of the frequency for representing the results of wavelet analysis. The translation parameter τ specifies location of the wavelet in time, by changing τ wavelet can be shifted over the signal. The elements in XWT (τ , s) are called wavelet coefficients, each wavelet coefficient is associated to a scale (frequency) and a point in the time domain. The WT also has an inverse transformation, as was the case for the FT and the STFT. The inverse continuous wavelet transformation (ICWT) is defined by equation (2).

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

---------- (2)

1.5 Discrete Wavelet Transform

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

1.6 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. The optimum wavelet is selected based on energy conservation properties in the approximation part of the wavelet coefficients as shown in [10]. In proposed work we have used Daubechies wavelet. Figure 1 shows Daubechies wavelet.

Figure 1: Daubechies Wavelet

1.7 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. We now introduce the concept of an accumulated distance matrix (ADM). The ADM contains the respective value in the local distance matrix plus the smallest neighboring accumulated distance. We can use this matrix 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.

2. BLOCK DIAGRAM

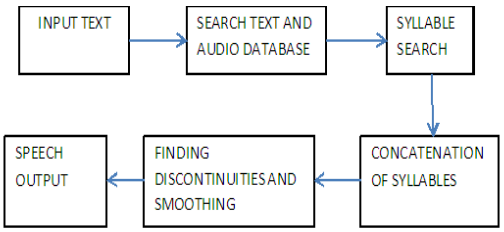

Figure 2: Block Diagram of System

Block Diagram explanation is given below

2.1 Text Processing and Audio Coding

Input to TTS synthesis system is Devnagari (Marathi) text. Encoding stage encodes input text into code string (CV structure).

2.2 Search for Word

Input text is searched in the text and audio database. Two databases are maintained viz. Audio database and Textual database. Audio database stores recorded sound files. Sound files are stored in .WAV format. The Textual database is required to search the index of the required word in the Audio database.

2.3 Search for Syllable

If word is not present in database, then system searches for syllable. After finding syllables they are given to the next stage for concatenation.

2.4 Concatenation of Syllable

Concatenations of words are made to generate proper-concatenated word and improper-concatenated words, (syllable positions taken into consideration) so as to compare the original word with the concatenated words and find out the difference. [9]

2.5 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 smoothing. [4] The parameters used are: 1) Power Spectral Density 2) Wavelet Transform 3) Dynamic Time Warping 4) Back-propagation

2.6 Speech Output

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

3. SYSTEM FLOWCHART

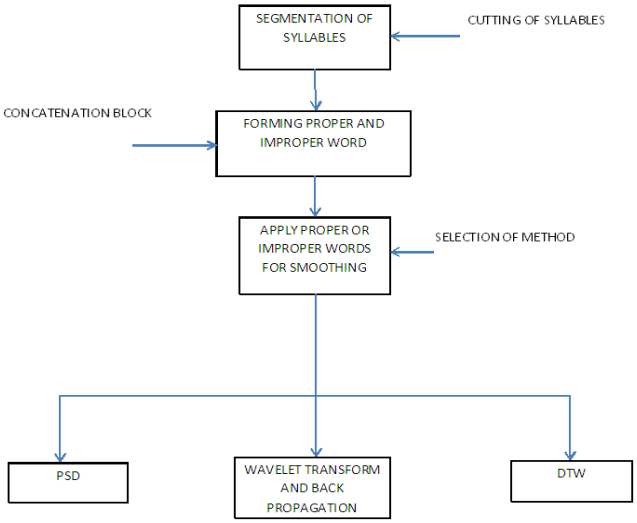

Figure 3: System flowchart

3.1 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

3.2 Power Spectral Density

PSD is a very useful tool if you want to identify oscillatory signals in your time series data and want to know their amplitude as shown in [2]. In proposed work PSD is used to find discontinuities in frequencies and for smoothing.

$$PSD = |X(f)|^2/N \quad \text{---------- (3)}$$

Where N is window size

3.3 Wavelet Transform

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.

$$\emptyset(x) = \sum_{k=-\infty}^{\infty} a_x \emptyset(Sx - k) \quad \text{---------- (4)}$$

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.

3.4 Dynamic Time Warping

In DTW, after calculating FFT for both original and proper (or improper), one matrix is assigned to original and other to proper word (or improper). 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)}

---------- (5)

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.

3.5 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.

3.5.1 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.

4. RESULTS

4.1 PSD Results

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

4.1.1 Results for Two Syllable words

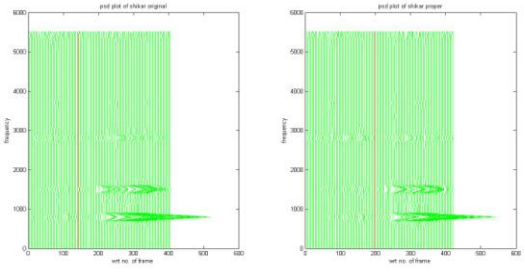

Figure 4: PSD plot of original Shikar and properly concatenated Shikar.

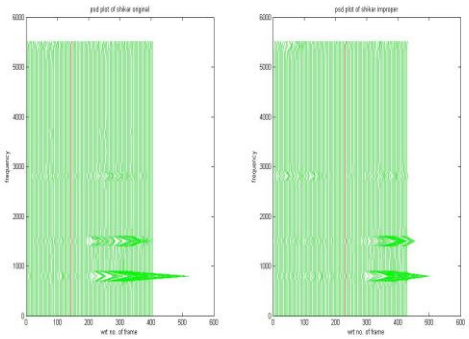

Figure 5: PSD plot of original 'Shikar' and improperly concatenated 'Shikar'.

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 original word. Thus properly concatenated word is more similar to original than improperly concatenated word.

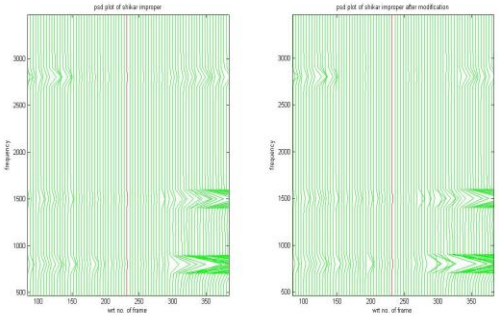

Figure 6: PSD plot of improper 'Shikar' and modified PSD plot of improper 'Shikar'.

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

4.1.2 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 1 shows the 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 in the proper and improper words. The mismatch is more near the concatenation frame which is shown in red color.[5]

4.1.3 Results for Three Syllable words

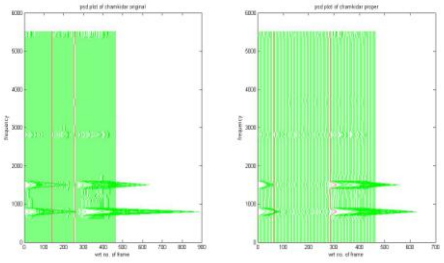

Figure 7: PSD plot of original 'Chamkidar' and properly concatenated 'Chamkidar'.

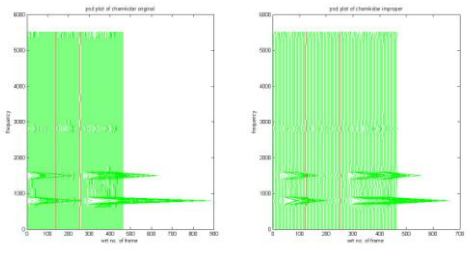

Figure 8: PSD plot of original 'Chamkidar' and improperly concatenated 'Chamkidar'

There are two cutting points in the figure 8 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'. It can be seen from the plots that there is large difference in the formants of improper and original word.

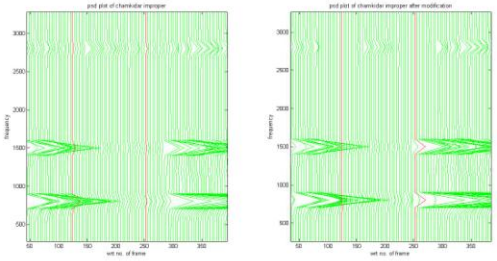

Figure 9: PSD plot of original 'Chamkidar' and improperly concatenated 'Chamkidar' after modification.

4.2 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.

Table 2 Numerical results for word 'Kamdharana'.

First Formant		Second Formant		Third Formant	
Original-Proper	Original-Improper	Original-Proper	Original-Improper	Original-Proper	Original-Improper
First Cutting Point					
36.27805	10.07811	2.891668	5.023969	3.145258	4.080231
33.66358	9.155186	3.055134	0.712101	3.943713	5.379192
13.00875	0.280157	4.292516	0.768871	7.061861	8.139648
4.06879	8.085539	0.125269	3.248089	5.745949	7.017595
7.865777	4.570448	0.568614	1.234643	5.765296	6.982857
10.7629	7.868121	0.580662	0.8266	5.405784	5.317275
2.993148	5.683566	1.335886	1.007553	0.932428	2.643646
16.99962	9.232243	16.70605	0.658192	3.917732	6.288512
18.59749	10.15083	21.18729	6.144337	9.186845	9.182986
18.33306	5.685015	23.47998	1.639247	4.563777	4.729203
26.60697	1.48352	17.76912	1.7026	3.930504	4.041851
20.80796	0.578751	2.587397	1.2637	0.480476	0.510656
28.01041	1.455568	8.628993	0.205696	0.214801	1.320191
25.38261	2.974617	6.325874	0.910257	0.577276	1.449041
44.03956	5.375221	2.475307	1.349288	0.015699	1.798456
26.51792	2.527094	4.063266	0.982963	0.155259	0.737049
19.0186	4.482236	19.5535	0.255813	1.005329	0.028725
45.90944	26.55739	28.4668	6.632104	0.237315	1.486242
68.33064	11.23094	33.54166	20.37203	2.736582	1.474859
77.31558	1.417367	3.357287	35.52404	0.785367	2.158093
114.6353	16.12562	11.04742	33.74665	6.692223	6.304684
Second Cutting Point					
33.92581	72.26906	9.371313	29.7966	13.60256	3.312138
14.59645	29.10085	20.90049	8.817026	11.3407	1.981676
35.08267	22.26339	27.2529	20.09064	3.551877	1.946423
43.86819	38.38172	38.74977	40.91644	0.410496	0.431085
14.89505	17.02978	44.2576	41.8463	4.350424	3.795476
3.774988	8.415386	37.15206	13.31233	2.710941	4.578499
15.44003	9.534411	10.76163	36.24272	0.108098	3.95355
9.843838	2.925325	54.55704	21.74958	2.023395	5.623964
23.05398	2.807055	34.40608	8.071655	1.286419	1.007893
9.398537	4.114244	20.93181	10.91756	6.549263	5.637051
0.60372	3.935164	5.579357	7.448092	5.986308	6.834373
4.759237	6.782752	5.743766	8.791217	6.371411	5.493872
3.685446	5.129344	0.730931	2.246993	4.381086	3.357424
5.150725	7.780251	4.627136	4.576095	2.322764	3.003323
2.149388	0.149993	8.16507	7.919301	3.986215	3.184112
3.357263	1.749767	0.3144	12.36594	2.357354	3.083138
27.24548	0.838936	45.37261	66.76282	2.180648	0.80172
26.86963	4.85197	104.9637	100.2681	10.02977	9.307106
10.3362	14.38211	79.06561	66.42473	0.734541	1.094017
13.62374	30.22268	10.87184	17.3922	4.158423	9.545652
58.4716	52.32296	68.618	108.3607	13.67478	11.52841

4.2.1 Results for Four Syllable Word

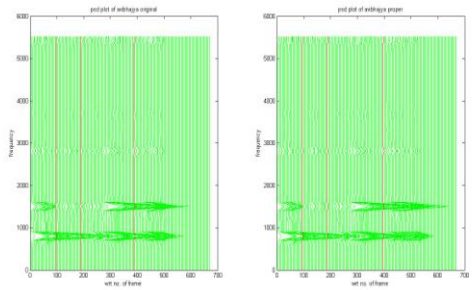

Figure 10: PSD plot of original 'Avibhajya' and properly concatenated 'Avibhajya'

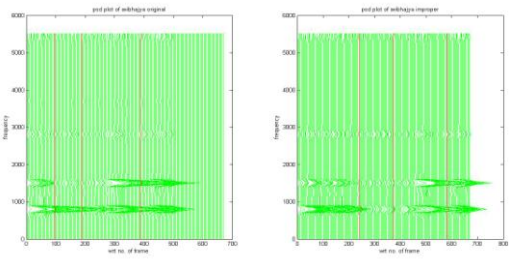

Figure 11: PSD plot of original 'Avibhajya' and improperly concatenated 'Avibhajya'

In figure 10 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'.

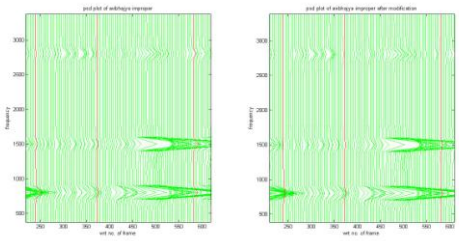

Figure 12: PSD plot of original 'Avibhajya' and improperly concatenated 'Avibhajya' after modification

4.3 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. One can get similar results like table 1 or table 2 for 4 syllable or polysyllable words.

From the tables 1, 2 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. [6][7]

5. 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 coefficient.

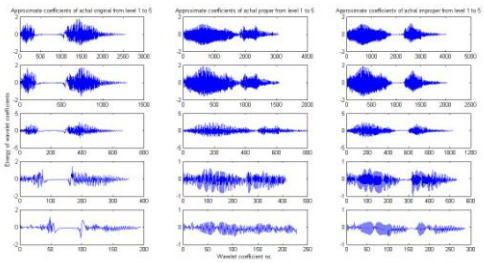

Figure 13: Approximate wavelet coefficients of 'Achal' from level 1 to 5

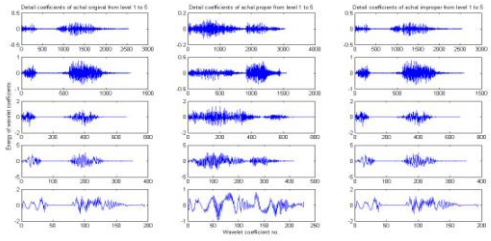

Figure 14: Detail wavelet coefficients of 'Achal' from level 1 to 5

It can be seen from the figure that there is mismatch in the wavelet coefficients of proper and improper word. 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.

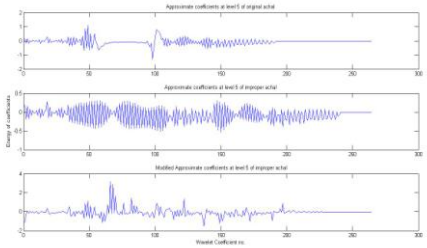

Figure 15: Approximate wavelet coefficients of original, improper and modified improper 'Achal'

It can be seen from the figure 15 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.

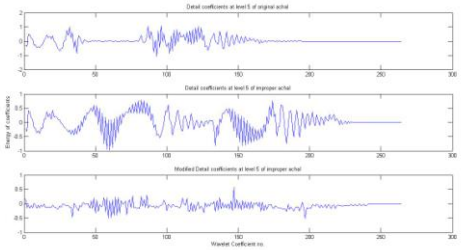

Figure 16: Detail wavelet coefficients of original, improper and modified improper 'Achal'

It can be seen from the figure 16 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. % error is calculated as follows:

1) Average is calculated for original, improper and modified improper coefficients. Let's denote this average by Ao, Ai and Ami for original, improper and modified improper coefficients respectively.
2) Then % error between original and improper coefficients is calculated as

$$\% \text{ error o-i} = (\text{absolute}(Ao-Ai))/Ao*100 \quad \text{---------- (6)}$$

3) And % error between original and modified improper coefficients is calculated as

$$\% \text{ error o-im} = (\text{absolute}(Ao-Ami))/Ao*100 \quad \text{---------- (7)}$$

Table 3 and 4 shows % error calculation.

Table 3: Numerical results of Wavelet and Back-propagation for approximate coefficients

Word	Percentage Error Orig-Imp	Percentage Error Orig-M Imp
Achal	38.41	13.86
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.

Table 4: Numerical results of Wavelet and Back-propagation for Detail coefficients

Word	Percentage Error Orig-Imp	Percentage Error Orig-M Imp
Achal	55.43	32.89
Geetkar	36.48	22.37
Kamgar	45.22	63.36
Ramdev	68.0	18.51
Marekari	54.82	34.72
Maydesh	33.0	16.19
Savdhan	40.79	11.12
Varkari	37.99	17.90
Upay	51.36	34.22
Vinayak	52.74	10.62

6. 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. The cross-correlation is obtained directly by the inbuilt MATLAB function. Let Cop and Coi be the cross-correlation between original and proper and original and improper. Then formula used for calculating cross-correlation is

$$\sum_{i=1}^{N} Cop\ (i)/N \text{ ---------- (8)}$$

Following figures and table shows results of correlation.

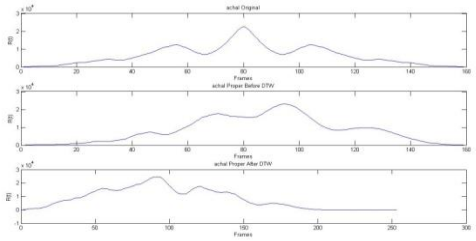

Fig 17: Original and Proper Anchal before and after applying DTW

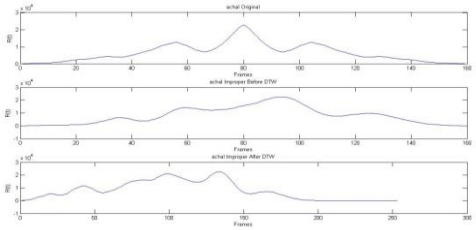

Fig 18: Original and Improper Anchal before and after applying DTW

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 5: 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	$2.3056*10^{6}$	$2.4586*10^{4}$	$2.2525*10^{4}$	$2.2587*10^{4}$
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	$7.8091*10^{3}$	$9.1944*10^{3}$	$6.4715*10^{3}$	$7.0319*10^{3}$

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.

7. CONCLUSION

All three methods, PSD, Wavelet and DTW help to estimate spectral mismatch (spectral artifacts) and for reduction of such mismatch. 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. The red line is concatenation point. 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 frequency domain parameter like DFW.

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 ofspontaneous speech into syllable like units", -T. Nagarajan, H.A. Murthy, I.I.T. Madras.

[10] "A Study on the Performance of Wavelet Packets for Spectral Analysis" M.K. Lakshmanan et.al, IRCTR, Dept of Electrical Engg, Delft University, Netherlands.