Language Independent Automated Segmentation of Speech using Bach scale filter-banks

G. Ananthakrishnan[1], H.G. Ranjani[2], A.G. Ramakrishnan[3]
[1]Dept. of Electrical Engineering Dept, Indian Institute of Science,
Bangalore – 560012, INDIA, ananthg@ee.iisc.ernet.in
[2]Dept. of Electrical Engineering Dept, Indian Institute of Science,
Bangalore – 560012, INDIA, ranjani@ee.iisc.ernet.in
[3]Dept. of Electrical Engineering Dept, Indian Institute of Science,
Bangalore – 560012, INDIA, ramkiag@ee.iisc.ernet.in

Abstract

This correspondence describes a method for automated segmentation of speech. The method proposed in this paper uses a specially designed filter-bank called Bach filter-bank which makes use of 'music' related perception criteria. The speech signal is treated as continuously time varying signal as against a short time stationary model. A comparative study has been made of the performances using Mel, Bark and Bach scale filter banks. The preliminary results show up to 80 % matches within 20 ms of the manually segmented data, without any information of the content of the text and without any language dependence. The Bach filters are seen to marginally outperform the other filters.

1. Introduction

For the purpose of speech synthesis, the corpus needs to be segmented into phonetic units. Manual segmentation is often tedious and time consuming. This calls for automated methods for doing the same. This correspondence describes one such automatic method.

The earliest attempts at automated segmentation were using the spectrogram of the signal and counting the number of zero-crossings in a region of speech.[1,2] Van Hemert [3] used the intra frame correlation measure between spectral features to obtain the segments. Statistical modeling (AR, ARMA) [4] has also been used. HMM based automated phonetic segmentation [5] requires a great amount of training, but provides excellent results. The most popularly used feature vector based methods are the Spectral Transition Measure (STM) and the Maximum Likelihood (ML) segmentation methods [6]. Another method using Average levels crossings has been suggested called (A-LCR) by [7].
HMM based segmentation gives the best results but needs high amount of training data, while the other methods mentioned do not require training.

It has been shown that Mel Frequency scale [8] and Bark scales are based on perceptual measures of acoustics. However not much has been explored of speech signals using the music perception based 'Bach' scale. This paper endeavors to show that the 'Bach' scale filter-bank shows a slightly better performance as compared to the 'Mel' and the 'Bark' scales.

2. Problem Formulation

A. *Design of Filter-bank*

The inspiration for the Bach scale is obtained from music, where there are 12 semi-tones in an octave. Each of the semitones is related to the next one by roughly a ratio of $2^{(1/12)}$. This ratio was initially discovered by the great musician of the 18th century, J.S. Bach [9,10]. This magical number of $2^{(1/12)}$ holds true for almost all genres of music and relates to some natural perceptual phenomenon. The first problem is, to design a filter bank corresponding to this scale.

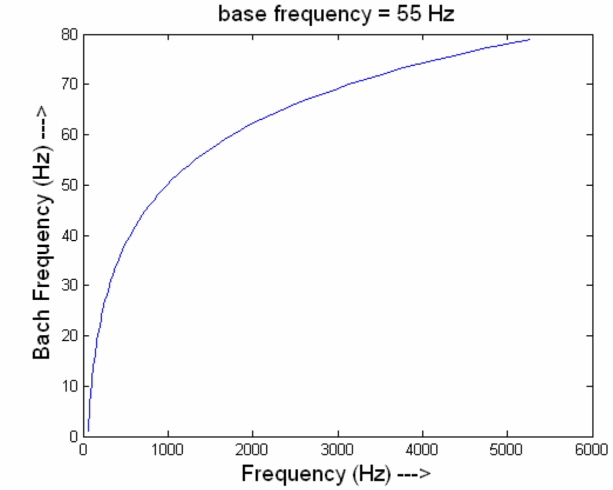

Fig. 1: The 'Bach' scale

B. *Obtaining feature vectors*

The most common methods of analyzing the time-varyng speech signal is by treating it as short-time stationary. However, this correspondence considers the speech signal as time varying.

The speech signal is filtered by each of the filters in the filter bank. So for 'N' filter-banks, we obtain 'N' filtered versions of the signal. Thus for every time instant we obtain

‘N’ feature vectors corresponding the output energies of the filters.

C. Detecting phoneme segment boundaries

Speech is considered as a sequence of quasi-stationary units called phones. Segmentation should ideally segregate the signal into such quasi-stationary units. However due to co-articulation effects the boundaries are not clearly defined. The effect of a phone is observed within regions of the preceding and succeeding phones.

In order to solve this problem, we need to find a region with minimum influence of the phones on either side of a boundary. The method employed is thus. Consider a 100 ms window. Divide this window into two equal halves. Obtain the Euclidian distance between the means of the feature vectors of the two halves of the window. This distance measure can thus be called the ‘Mean Euclidian Distance’ (MED). This way, for every instant if time we can find the corresponding MED value. The phone segments are the regions between two consecutive peaks occurring in this MED function. This ensures that in presence of co-articulation, the boundary is obtained where there is minimum presence of either of the neighboring phones.

3. Problem Solution

A. Design of Filter-bank

The centre frequency (f_c) of the ‘nth’ filter-bank is obtained by the equation,

$$f_c(n) = \text{base}*(2^{n/12})/F_s \qquad (1)$$

where ‘F_s’ is the sampling frequency and ‘base’ is the starting frequency of the filter-bank. ‘n’ is also equivalent to the relative ‘Bach’ frequency

The maximum possible centre frequency of the filter-bank (MaxFreq) is calculated by

$$\text{MaxFreq} = 12*\log_2(F_s/(2*\text{base})) \qquad (2)$$

There are two ways of formulating the bandwidth (f_b) of the ‘nth’ filter

$$f_b(n) = \text{base}*(2^{(n+1)/12} - 2^{(n-1)/12})/(Fs*2) \qquad (3)$$

Or

$$f_b(n) = \text{base}*(2^{(2^{(n-1)*\log_2((\text{MaxFreq}-12)/12)/\text{MaxFreq}} -1)})/F_s \qquad (4)$$

The bandwidth formulation given by (3) gives a linear change in cut-off frequency with respect to central frequency. The second formulation (4) gives a non-linear change in cut-off frequency with respect to the central frequency shown in Fig. 2

The number of filter coefficients used to generate the ‘nth’ window is determined by

$$N(n) = 2*ceil(1/\, f_b(n)) \qquad (5)$$

The filters designed are lag-windows designed by the standard Blackman-Tukey spectral estimation method. [11] The set of filter coefficients obtained, is the eigenvector associated to the maximum Eigen value of the matrix with elements

$$\gamma_{m,n} = \beta * signum((m-n)*\beta*\Pi) \qquad (6)$$

where 2*β is the band-width in radians/sec and

$$\begin{aligned} signum(x) &= \sin(x)/x \qquad \{x != 0\} \\ &= 1 \qquad \{x = 0\} \end{aligned} \qquad (7)$$

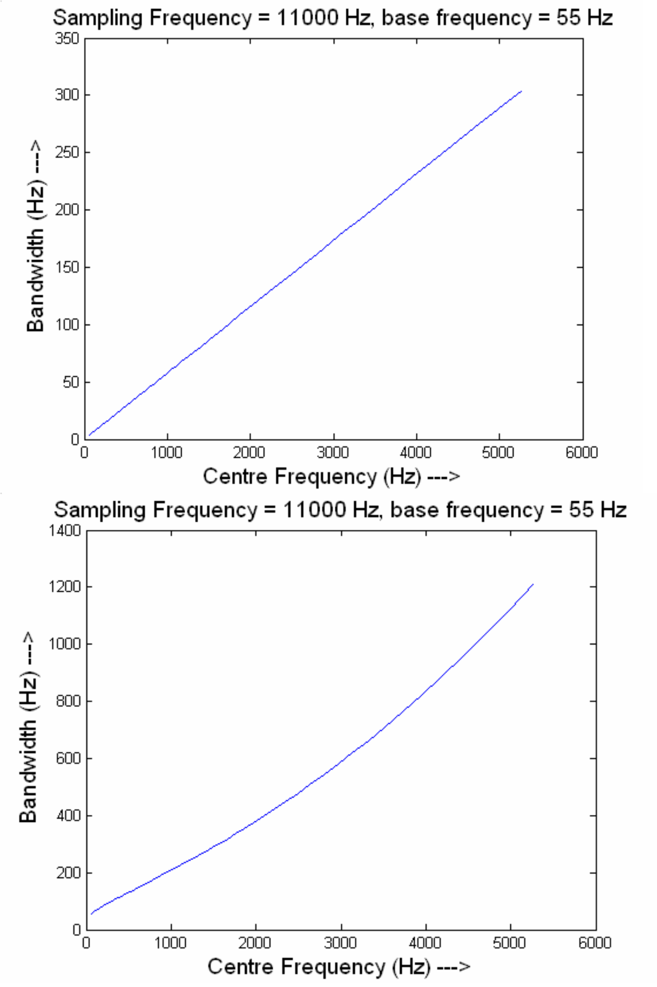

Fig. 2: Bandwidth of the filters as against the centre frequency in Hz Linear case (3), (b) Non-Linear case (4)

The filter coefficients are real, symmetric and finite, so the phase responses are linear. The magnitude response of the set of filters is shown in Fig. 5.

B. *Obtaining Feature vectors*

The number of feature vectors obtained depends on the ‘base’ frequency and the sampling frequency of the speech signal. The base frequency is a parameter variable which can be determined by the type of speech data. Any frequency between 50 and 80 Hz gives a good performance. As an example base = 50 Hz and F_s = 11000 Hz give ‘81’ feature vectors. The ‘n^{th}’ feature vector for speech sample ‘k’ is represented by $F_k(n)$ or $F_n(k)$.

$$F_k(n) = F_n(k) = abs\left(h_n(k) \otimes s(k)\right) \qquad (8)$$

where s(k) is the input speech signal, $h_n(k)$ is the band-pass filter designed around centre frequency ‘n’. The $\otimes$ symbol represents linear filtering. The feature vectors $|F_k(n)|_{k=1:T}$ or $|F_n(k)|_{n=1:M}$ are the two ways of the 2-D representation of the signal s(k).

The filter-bank is only an analysis filter-bank and not a perfect reconstruction one. Since the number of coefficients for the filter is inversely proportional to the bandwidth of the filters, we get better time resolution in higher frequencies and better frequency resolution at lower frequencies.

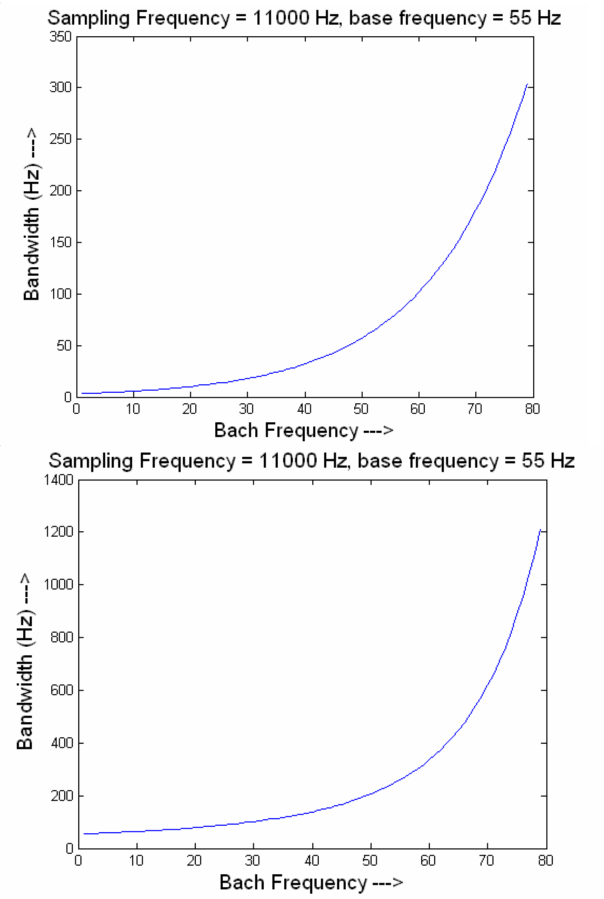

Fig. 3: Bandwidth of the filters as against the centre frequency in Bach
(a) Linear case (3), (b) Non-Linear case (4)

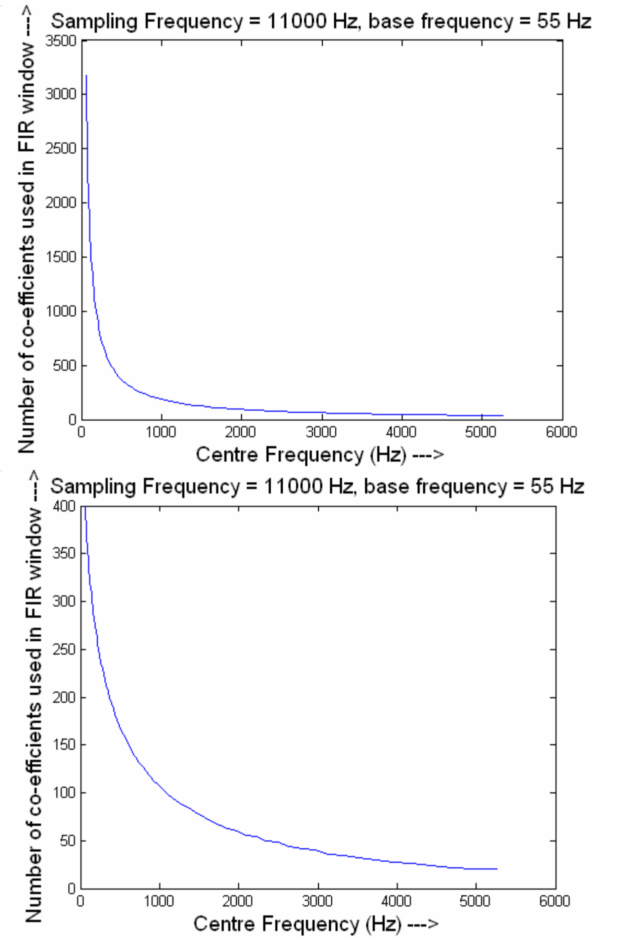

Fig. 4: The number of filter coefficients used to design the FIR filter depending on the centre frequency
(a) Linear case (3), (b) Non-Linear case (4)

C. *Obtaining the Phone Segments*

For ‘k^{th}’ speech sample the ‘Mean Euclidian distance’ (MED) is calculated as follows

$$M1(n) = \frac{\sum_{i=k-W}^{k} F_i(n)}{W}$$

$$M2(n) = \frac{\sum_{i=k}^{k+W} F_i(n)}{W} \qquad (9)$$

$$MED(k) = \| M1 - M2 \| \qquad (10)$$

Where ‘W’ is the length of the region under consideration. W is a parameter which should be set to around twice the average phone duration. Since information about the language or the sequence of phonemes is assumed not to be available W is set to a constant value of 100 ms. If such information as the phoneme sequence is available, then it could be incorporated in deciding the value of ‘W’, which then could be a variable quantity.

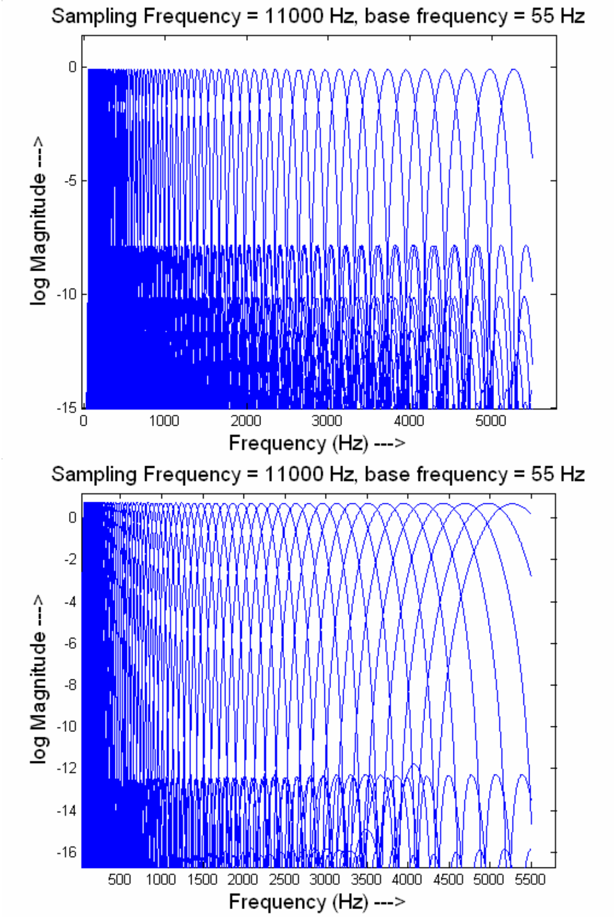

Fig. 5: The Bach scale Filter-bank
Linear case (3), (b) Non-Linear case (4)

We now know that the MED function gives an indication of the difference in spectral properties on either side of the 'kth' sample within the specified region of consideration. The segment boundary is thus attributed to the point of maximum difference between the two sides of a sample of speech.

To convert the MED to segment boundaries we have to just detect the peaks of the MED waveform. Here the intensity of the peak is not relevant for segmentation. The existence of the peak itself is of importance. However due to modulations in the MED function, peak detection in itself poses a problem.

D. Leading Slope Stressed function (LSSF)

A method to determine how important a peak in the MED waveform is by finding the LSSF.

The LSSF at 'k' for a region 'R' is given by

$$[m,i] = \min(MED((k-R):k))$$
$$LSSF(k) = (MED(k) - m)/(k - i) \qquad (11)$$

Here another parameter is of importance, namely the inertia of the system. The parameter 'R' is determined by how quickly the phones change. So 'R' should be selected such that it is less than the shortest possible phone length and larger than temporal variations within the phone. It is also determined by the value of 'W'. Typically it is set between 10 and 20 ms.

The LSSF now gives a waveform that has peaks, whose amplitude depends on the importance of the peak under the given context. The larger the difference between adjacent phones, the higher is the amplitude of the peaks of LSSF.

The actual peak detection is achieved by using this simple method.

$$\begin{aligned} &if \\ &LSSF(k) = \max(LSSF(k - R/2 : k + R/2) \\ &then \qquad Peak(k) = 1 \\ &else \qquad Peak(k) = 0 \end{aligned} \qquad (12)$$

For every 'k' for which Peak(k) = 1, is considered a segment boundary. The LSSF increases the performance of the segmentation by a considerable amount and also enables us to put a threshold of picking the significant peaks. LSSF gives the maximum slope in region 'R'. Thus it peaks at places where maximum change in the MED function occurs. The selection of length of 'R' is also critical. If set too small, Lot of false boundaries are picked up and if set too large too less boundaries are picked up.

4. Results and Conclusion

The quality of segmentation is evaluated by comparing the output of the automated segmentation algorithm with manually segmented databases. If an automated segment boundary falls within 20 ms of a manually segmented boundary, then it is considered to be a 'Matched phone boundary' (MPB).

If more than one automated segment boundary falls within ±20 ms of a manual boundary or no manual boundary is found within ±40 ms of an automated boundary then such boundaries are considered to be 'insertions'(Ins). On the other hand if no automated boundary is found within ±40 ms of a manual boundary, then it is considered as a 'deletion' (Del). The percentage accuracy is calculated as

$$\frac{\text{(No. of correct phone boundaries)*100}}{\text{(Total no. of manually segmented phone boundaries)}} \qquad (13)$$

The results are obtained for 100 sentences of English data from the (F_s = 16000 Hz) TIMIT database for both male and female speakers. The data has an SNR of 36 dB. 100 sentences of Hindi and Tamil data have also been used. This data has a sampling frequency of 44.1 KHz and an SNR of 30 dB. The data available for Tamil and Hindi are only that of a male voice.

From Table 1, we can see that the 'Bach linear' and the 'Bach non-linear' scales perform comparably if not marginally better than the 'Mel' or 'Bark' scales. We can however see a reasonably significant difference in the number of false inserted phone boundaries between the 'Mel' and 'Bark' scales as against the 'Bach' scales. However it can be noted that the number of deletions of the boundaries are higher in case of the Bach (Lin) case.

TABLE 1 – COMPARISON BETWEEN THE VARIOUS FILTER-BANK SCALES FOR TIMIT DATABASE

Filter- Bank type	%MPB	%Del	%Ins
Mel	78.1	16.5	68.1
Bark	78.1	17.9	50.1
Bach (Lin)	82.5	22.3	18.9
Bach (Non-Lin)	79.3	17.4	20.4

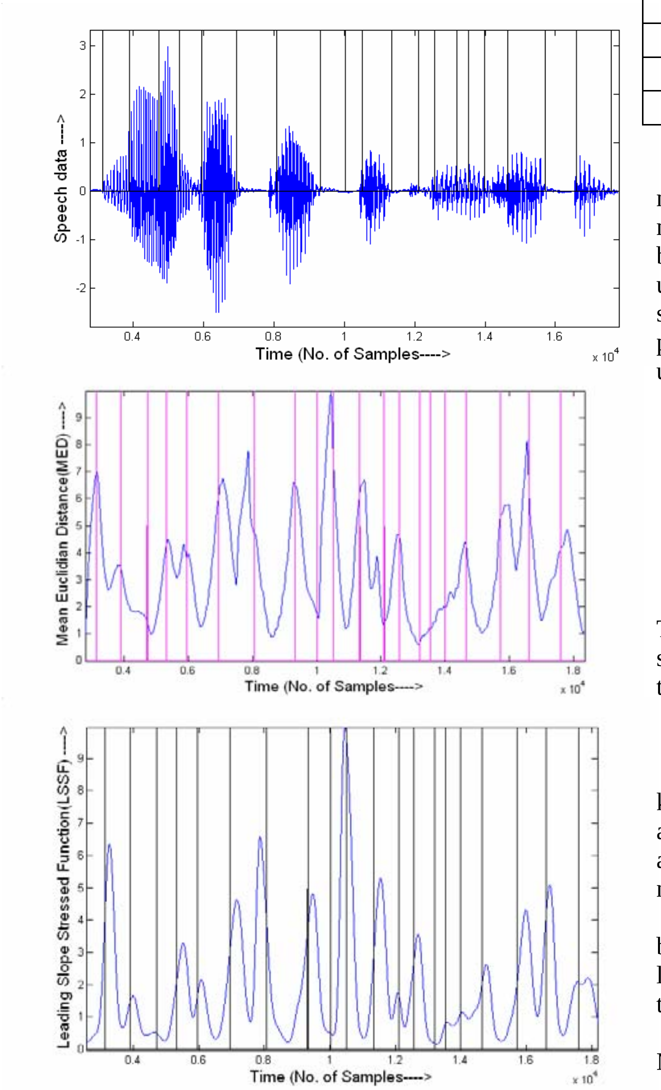

Fig. 6: (a) The speech signal with manual boundaries marked (b) The MED function plot along with the manual boundaries (c) The LSSF function plot along with the manual boundaries

Two spectral domain methods and one time domain method has been used on comparative basis to study the proposed method.

1. ML Segmentation using MFCC with a symmetric lifter $(1 + A\sin 1/2(\pi n/L))$ (A = 4, L is the MFCC dimension = 16) [6].

2. Spectral Transition measure (STM) using feature vector and lifter combination.

3. Average level crossing rate method (A-LCR) as described in [7] using non-uniform level allocation.

TABLE 2 – COMPARISON BETWEEN VARIOUS SEGMENTATION METHODS ON THE TIMIT DATABASE

Segmentation Method Used	%MPB	%Del	%Ins
ML[6]	80.8	19.2	18.8
STM[6]	70.1	29.9	25.2
A-LCR [7]	79.8	20.2	24.2
LSSF (Bach Lin)	82.5	22.3	18.9

Table 2 compares the performances of the proposed method and the other standard methods. The proposed method does marginally better in terms of ‘matched phoneme boundary’ (MPB) percentage. However the standard methods use information such as the number of phones and location of silences, in order to obtain the correct phone boundaries. The proposed method using LSSF gets similar results without using such information.

TABLE 3 – COMPARISON THE PERFORMANCE OF LSSF USING BACH NON-LINEAR FILTER-BANK FOR VARIOUS LANGUAGES

Language	%MPB	%Del	%Ins
English	82.5	22.3	18.9
Hindi	79.6	10.7	32.5
Tamil	72.1	15.3	23.7

Table 3 shows that the proposed method is language and speaker independent, showing comparable results for all the three languages.

5. FUTURE WORK

Future work can be carried out in terms of incorporating knowledge of the phones and linguistic knowledge like average duration of the phones etc. Noise robustness of the algorithm can also be tested and special considerations for noise robustness can included in the algorithm.

Another interesting area could be use of the ‘Bach’ filter-bank in areas like speech recognition and defining features like the ‘Bach Frequency Cepstral Coefficients’ and compare their performance with Mel and Bark scales.

We could define several distance measures instead of the MED defined in this paper and evaluate comparative results.

REFERENCES

[1] D.R. Reddy, “Segmentation of Speech Sounds”, J.Acoust.Soc.Am.-1966, Vol. 40, No. 2, pp-307-312

[2] James R. Glass and Victor W. Zue, “Multi Level Acoustic Segmentation of Continuous Speech”, Proc. of ICASSP- 1988, pp: 429-432.

[3] Jan P. van Hemert, “Automatic Segmentation of Speech”, IEEE Trans. on Signal Proc., Vol. 39, No. 4, April 1991, pp-1008-1012.

[4] R. Andre-Obrecht, “Automatic Segmentation of Continuous Speech Signals”, Proc. ICASSP-Tokyo, 1986, pp-2275- 2278.

[5] D.T. Toledano, L.A. Hernandez Gomez and L.V. Grande, “Automatic Phonetic Segmentation”, IEEE Trans. Speech and Audio Proc., Vol 11, No. 6, Nov. 2003, pp 617-625

[6] T. Svendsen and F.K. Soong, “On the Automatic Segmentation of Speech Signals”, Proc. ICASSP-Dallas, 1987, pp: 77-80.

[7] Anindya Sarkar and T.V. Srinivas, “Automatic Speech Segmentation Using Average Level Crossing Rate Information”, Proc ICASSP, 2005, pp: I-397 to I-400

[8] L. Rabiner and B. H. Juang, “Fundamentals of Speech Recognition”, Pearson education Press, 1993 edition (AT&T)

[9] N. Slonimsky, Thesaurus of Scales and Melodic Patterns (1947);

[10] C. Sachs, The Wellsprings of Music (1965).

[11] Petre Stoica and Randolph L. Moses, “Introduction to Spectral Analysis”, Prentice Hall, pp 46-48

[12] A.K.V. SaiJayram, V.Ramasubramanian and T.V. Sreenivas, “Robust parameters for automatic segmentation of speech”, Proc. ICASSP-May,2002, pp-1-513-1-516