COMPARATIVE STUDY OF FILTER-BANK MEAN-ENERGY DISTANCE FOR AUTOMATED SEGMENTATION OF SPEECH SIGNALS

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

ABSTRACT

This correspondence describes a method of automated segmentation of speech assuming the signal is continuously time varying rather than the traditional short time stationary model. It has been shown that this representation gives comparable if not marginally better results than the other techniques for automated segmentation. A formulation of the 'Bach' (music semi-tonal) frequency scale filter-bank is proposed. A comparative study has been made of the performances using Mel, Bark and Bach scale filter banks considering this model. 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. 'Bach' filters are seen to marginally outperform the other filters.

Index Terms **-** Segmentation, Filter-Banks, Mean Euclidian Distance, time varying model

1. INTRODUCTION

It has been shown that the Mel scale and the Critical band rate (Bark) scales [1] are based on perceptual properties of the human auditory system. The inspiration for the 'Bach' scale is obtained from music. In music, every octave contains 12 semi-tones. Each of the semitones is related to the next one by roughly a ratio of $2^{(1/12)}$. This ratio was discovered by the great musician of the 18th century, J.S. Bach [2, 3]. This number of $2^{(1/12)}$ holds true for almost all genres of music and relates to some natural perceptual phenomenon. The Mel and Bark scales have been shown to provide robust results for speech segmentation and recognition. However not much has been explored of speech signals using the music perception based 'Bach' scale.

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.

Automated segmentation has been attempted by counting the number of level-crossings in a region of speech [4], using the intra frame correlation measure between spectral features to obtain the segments namely the Spectral Transition method (STM) and the Maximum Likelihood (ML) [5], statistical modeling (AR, ARMA) [6,7] and HMM based methods [8]. HMM based segmentation gives the best results but needs high amount of training data, while the other methods mentioned do not require training.

2. PROBLEM FORMULATION

2.1. Time varying representation

The most common method of analyzing the time-varying speech signal has been by treating it as short-time stationary. However, this correspondence considers the speech signal as time varying. The speech signal is filtered by a bank of 'M' band-pass filters each shifted in frequency by a fixed factor. So we have 'M' filtered versions of the same speech signal. Consider the 'n^{th}' such version of the signal. The energy around the 'n^{th}' frequency component of the signal around a time instant 'k' will be equal to the 'k^{th}' output energy of the 'n^{th}' filter-bank.

$$F_k(n) = F_n(k) = abs\left(h_n(k) \otimes s(k)\right) \qquad (1)$$

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 first filter is centered around a 'base' frequency (any base freq between 50 to 80 Hz results in a good performance). The filter-bank is only an analysis filter-bank and not a perfect reconstruction one.

2.2. Comparison of scales

Filter-banks have been designed for 4 auditory scales namely the Mel or Radio scale, the critical band rate scale (CRB), equivalent rectangular band (ERB) rate scale and

the Bach scale. The first filter for all the banks is shifted by a 'base' frequency.

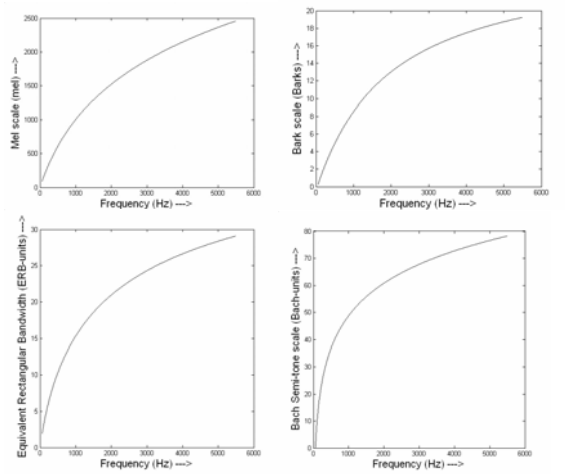

Fig. 1 – (a) The 'Mel' scale, (b) The Critical Band rate scale (Bark), (c) The Equivalent rectangular band scale (ERB), (d) The 'Bach' scale

The approximation to the experimental data for Mel scale is given by Beranek [9],

For Mel scale

$$m(f) = 1127 * \ln\left(1 + \frac{f}{700}\right) \quad (2)$$

Critical Band rate (Bark) scale [9]

$$z(f) = \frac{26.81}{1 + \frac{1960}{f}} - 0.53 \quad (3)$$

For Equivalent rectangular Band rate scale [10]

$$e(f) = 11.17 * \ln\left(\frac{f + 312}{f + 14675}\right) + 43.0 \quad (4)$$

For 'Bach' scale

$$b(f) = 12 * \log_2\left(\frac{f}{base}\right) \quad (5)$$

The Bach scale is only a relative scale and depends on the 'base' frequency.

Assuming 12 filters per octave corresponding to 12 semitones in the Bach scale, the maximum number of filters 'M' is calculated by.

$$M = 12 \log_2\left(\frac{F_s}{2 * base}\right) \quad (6)$$

where 'F_s' is the sampling frequency. For a good comparison, filter-banks with 'M' filters are used for all the other scales. The centre frequencies ($f_c(n)$) of the filters are obtained by uniformly sampling by 'M' in their respective frequency scales. 's' is the shift in the frequency for every next filter. So the centre frequency of the 'n^{th}' filter is given by s*n.

The bandwidths for the scales are calculated from the approximations to the experimental values.

For Mel [9]

$$B_{mel}(f) = \frac{700 * \left(e^{\frac{m(f)+s}{1127}} - e^{\frac{m(f)-s}{1127}} - 2\right)}{2} \quad (7)$$

Critical Bandwidth [9]

$$B_z(f) = \frac{52548}{z(f)^2 - 52.56 * z(f) + 690.39} \quad (8)$$

Equivalent rectangular band rate scale [10]

$$B_e(f) = 6.23 * 10^{-6} f^2 + 9.339 * 10^{-2} f + 28.52 \quad (9)$$

There are two ways of formulating the bandwidth of the 'Bach' scale filter :-

with a linear change with respect to central frequency

$$B_{bach}(f) = base * \left(2^{(b(f)+1)/12} - 2^{(b(f)-1)/12}\right) \quad (10)$$

Or with an exponential change with respect to central frequency

$$B_{bach}(f) = base * 2^{(b(f)-1) * \log_2\left(\frac{M-12}{12M}\right)} - 1 \quad (11)$$

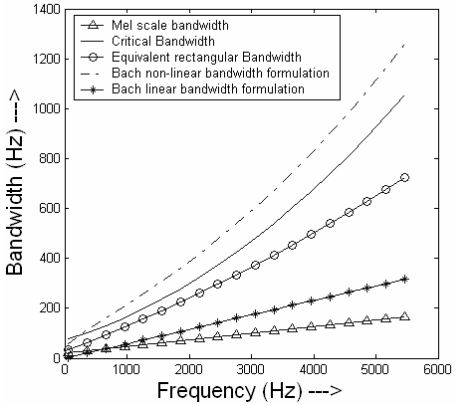

Fig. 2 – Bandwidth of the filters as against the centre frequency (Hz) for the (a) Mel scale (b) Critical Bandwidth (c) Equivalent rectangular Bandwidth (d) 'Bach' Non-linear scale, (e) "Bach' linear scale

The number of filter coefficients used to generate the 'n^{th}' filter is determined by

$$N(n) = 2 * ceil(1 / f_b(n)) \quad (12)$$

We thus see that the time resolution is poor for lower frequencies but better for higher frequencies. So we get the paradoxical ability to get better time resolution for higher frequencies and better frequency resolution for lower frequencies.

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

$$\gamma_{m,n} = \beta * signum((m-n) * \beta * \Pi) \quad (13)$$

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

$$signum(x) = \sin(x)/x \qquad \{x != 0\} \quad (14)$$

$$= 1 \qquad \{x = 0\}$$

The filter coefficients are real, symmetric and finite, so the phase responses are linear. The magnitude responses of the set of filters constructed by the 'Bach' non-linear scale are shown in Fig. 3.

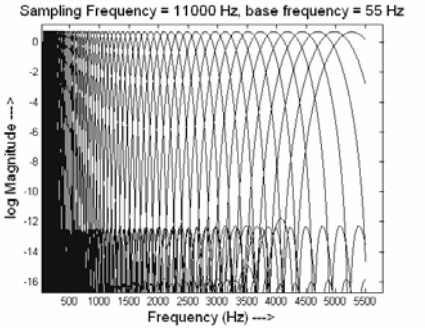

Fig. 3 – The Bach scale Filter-bank Non-linear case

3. 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

For 'kth' speech sample, the 'Mean Energy distance' (MED) is calculated as follows

$$M1(n) = \frac{\sum_{i=k-W}^{k} F_i(n)}{W} \qquad M2(n) = \frac{\sum_{i=k}^{k+W} F_i(n)}{W} \quad (15)$$

$$MED(k) = \| M1 - M2 \| \quad (16)$$

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 50 ms. If such information as the phoneme sequence is available, then it could be incorporated in deciding the value of 'W', which then would be a variable quantity

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, which corresponds to the peak in the MED function.

Here the intensity of the peak is not relevant for segmentation. However due to temporal modulations in the MED function, peak detection in itself poses a problem.

A function called Leading Slope Stressed function (LSSF) can determine how important a peak in the MED waveform is.

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

$$[m, i] = \min(MED(k-R:k))$$
$$LSSF(k) = (MED(k) - m)/(k - i) \quad (17)$$

The parameter 'R' or the inertia of the speech production mechanism 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. 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 this simple method.

$$\text{if } LSSF(k) = \max(LSSF((k-R/2):(k+R/2)) \quad (18)$$

Then *Peak(k)* = 1

Else *Peak(k)* = 0

Every 'k' for which Peak(k) = 1, is considered a segment boundary

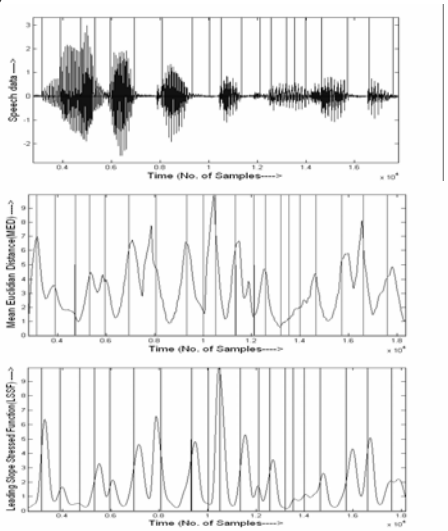

Fig. 4 – (a) The speech signal with manual boundaries, (b) The MED function plot along with the manual boundaries, (c) The LSSF function plot along with the manual boundaries

4. RESULTS AND CONCLUSION

The quality of automated segmentation is evaluated by comparing the output 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 unique automated boundary is found within ±40 ms of a manual boundary, then it is considered as a 'deletion' (Del).

The results are obtained for 100 sentences of English data from the (Fs = 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 segmented. The latter 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.

The following methods have 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) [12].
2. Spectral Transition measure (STM) using feature vector and lifter combination.
3. Average level crossing rate method (A-LCR) as described in [4] using non-uniform level allocation.

Table 1 – Comparison of performances between various segmentation methods on the TIMIT database

Segmentation Method	%MPB	%Del	%Ins
ML[5]	80.8	19.2	18.8
STM[5]	70.1	29.9	25.2
A-LCR [4]	79.8	20.2	24.2
LSSF (Bach Lin)	82.5	22.3	18.9

Table 2 – Comparison of performances between the various filter-bank scales on the TIMIT database

Filter- Bank type	%MPB	%Del	%Ins
Mel	78.1	16.5	68.1
Critical Band rate (Bark)	78.1	17.9	50.1
Equivalent Rectangular Band Rate (ERB)	76.3	18.9	52.4
Bach (Lin)	82.5	22.3	18.9
Bach (Non-Lin)	79.3	17.4	30.4

Table 3 – Comparison of the performance of LSSF using 'Bach' 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	76.1	15.3	23.7

Table 1 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 as information in order to obtain the correct phone boundaries. The proposed method using LSSF gets similar results without using such information.

From Table 2, 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 or inserted 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 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 . Noise robustness of the algorithm can also be tested and special considerations for noise robustness can be 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 the results.

REFERENCES

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

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

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

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

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

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

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

[8] 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

[9] LL Beranek, Acoustic Measurements, Wiley, New York, 1949), p.329.

[10] B.C.J. Moore and B.R. Glasberg (1983) "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns" J. Acoust. Soc. Am. 74: 750-753.

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

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