Music Signal Processing for Musicological Applications.

Iasonas Antonopoulos

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

Abstract. This thesis presents new methods and techniques for the study of inherent periodicities and rhythmic characteristics of polyphonic audio recordings. The above methodologies aim at the study of traditional music genres as is the case of Eastern Type traditions and Greek Traditional music. The above genres exhibit a great variety of rhythmic characteristics and are not investigated in depth in the existing literature. Toward this direction, the presented work takes into consideration the characteristics of the above music genres thus providing a tool for the study of Ethnomusicology.

1 Introduction

Content-based music analysis is a growing challenge in the context of Music Information Retrieval (MIR). One of the earliest and well studied topics of MIR is the automatic extraction and analysis of rhythmic features from audio recordings. The developed applications include tempo and music meter induction and beat tracking. Such features are justified by the temporal nature of music and the existence of inherent periodicities in music signals. Apart from the above, other applications based on rhythmic characteristics emerged over the years. Such are the retrieval of rhythmic similar recordings, the application of music summarization and other.

Throughout the years the methods exhibited have been focused on western music corpora and specific music styles. The latter resulted in the development of techniques that performed well when certain preconditions are satisfied. Such preconditions are the music meter type, the beat range and patterns of rhythm that consist of isochronously perceived tempo beats. These methodologies do not take into consideration less popular music genres as is the case with Eastern type traditions. In such traditions, a variety of music meters and beats and more free-form rhythmic patterns are very frequently encountered. The above poses a question for the Ethnomusicology studies that are in need of different approaches and techniques.

In this thesis, we present content-based techniques for the automatic extraction of inherent periodicities and rhythmic characteristics from an audio recording that can be applied to non-western music corpora. The applications

*Dissertation Advisor: Sergios Theodoridis, Professor

developed concern the induction of music meter and tempo,the automatic extraction of the two most similar segments of a music recording, that we call audio thumbnail, the music retrieval based on rhythmic similarity and a method for modeling and tracking of rhythmic patterns in an audio recording that may consist of uneven rhythmic components.

2 Rhythmic characteristics of Greek Traditional Dances

Rhythm and rhythmic patterns lie in the core of folk music, as is the case with Eastern traditional folk dances. This is also true for Greek traditional dance music [1], [2]. Unlike to the western music genres, a great variance of tempo ranges and music meters is exhibited. Tempo can range from $40 - 480$ *bpm* and music meters $\frac{2}{4}$, $\frac{3}{4}$, $\frac{4}{4}$, $\frac{5}{4}$, $\frac{7}{8}$, $\frac{9}{8}$, $\frac{12}{8}$, etc. A table of the studied corpus is exhibited in Table 1.

Table 1. Rhythmic tempo range, music meters and patterns studied.

3 Feature Extraction

3.1 Scale-based MFCCs

In order to proceed with the proposed applications, a feature extraction step takes place. In the feature extraction step, the audio recording is long-term segmented and for each long-term segments a short-term moving window analysis [3] takes place with overlapping Hamming windows of ~ 92 msec and overlap ~ 23 msec. For the short term analysis of each long term audio segment, we considered both energy and mel frequency cepstral coefficients (MFCCs) [4]. In addition to the standard MFCCs, which assumes equally spaced critical band filters in the mel scale, we propose a filter bank consisting of overlapping triangular filters, whose center frequencies align with the chromatic tones [5]. We will refer to this feature sequence as *scale-based MFCCs*, $\bar{c}$, of dimensions $M \times L$ feature sequence, where M is the number of short-term windows and L the number of MFCC filters used.

3.2 Event detection and Inter-Onset-Interval

An additional processing step takes place for the detection of music event beginnings and durations, for the application of modeling and tracking rhythmic patterns.

Let $stdMel(n)$, be the smoothed and normalized standard deviation of *scale-based MFCCs* and $dEner(n)$, $n = 1, \ldots, N$, the first derivative of signal energy for each frame, where N is the number of short-term frames. A peak picking

algorithm selects those maxima with frame index m for which $stdMel(m) > stdMel(k), \forall k \in [k_1, k_2]$ and m being the center of the $[k_1, k_2]$ interval. Let also i be the number of frame for which $dEner(i) > dEner(k), \forall k \in [k_1, k_2]$ with i being, also, the center of $[k_1, k_2]$. Our goal is to select those frames whose frames indices i, m coincide within a threshold value. For our applications this value was chosen to be equal to $0.1secs$. These frames are selected to indicate *onsets* and we choose the respective value of m to indicate the onsets. The value of $k_2 - k_1$ depends on the rhythmic components of the modeled rhythmic patterns and is usually set equal to the duration of the shortest rhythmic component of the pattern.

The physical meaning of these onsets is that they signal the beginning of an event, i.e., a *significant change* in terms of spectrum (*stdMel*) and energy (*dEner*). Each event will, therefore, have an onset and an associated time duration. Let, m_k, m_{k+1} be two consecutive selected onsets. Then $m_k < m_{k+1}$ and $m_{k+1} - m_k$ is the so-called inter-onset-interval (IOI). The feature sequence F will be given as input to the *HMM* modeling a rhythmic pattern, is formed by zeroing the *stdMel* of all frames, except those that correspond to onsets, i.e.,

$$F = \{O_{z_0}, a(m_1), O_{z_1}, a(m_2), \ldots, O_{z_{M-1}}, a(m_M), O_{z_M}\},$$

where O_{z_j} stands for z_j successive zeros. As a result, $a(m_j)$ is the amplitude of the j-th onset and O_{z_j} it's associated duration.

4 Tempo and music meter induction

In this thesis we propose an alternative method for the extraction of the tempo and music meter periodicities [5]. The proposed technique is based on Self Similarity Matrix analysis that was originally mentioned in [6]. Each long-term segment serves as the basis to generate a Self Similarity Matrix (SSM), using the Euclidean Distance metric. By its definition, the SSM is symmetric around the main diagonal and it therefore suffices to focus on its lower triangle. The diagonals of the SSM express the difference of the audio feature with different instances of itself and with the proper processing can reveal the pairs of correlating periodicities. If the mean of SSM diagonals is perceived as a function of the short-term step k then the following equation occurs:

$$S_1(k) = \frac{1}{M-k} \sum_{i=k}^{M} \left\| \bar{c}(i), \bar{c}(i-k) \right\|, k = 1, \ldots, M-1, \quad (1)$$

where k is the diagonal index and $\|.\|$ is the Euclidean distance function and $\bar{c}$ is the *scale based MFCCs*. Clearly, M-k is the length of the k-th diagonal. $S_l(k)$ is computed for all long-term segments, $l = 1, \ldots, L$. The plot of S_l as a function of k exhibits a number of minima and maxima. Since we employ the Euclidean Distance Metric the minima (valleys) indicate the similarity between feature instances. For the detection of the most significant periodicities (minima), we have developed an algorithm based on a dominance region criterion.

Specifically, a valley at lag m is considered to be *"dominant"* in a region, $[k_1, k_2]$, of lags, if it holds the lowest value in the region, i.e., if $S_l(m) < S_l(k), \forall k \in [k_1, k_2]$, where m is the center of $[k_1, k_2]$. The selected dominant minima will form pairs that comply with the studied correlated periodicities and music meters, i.e., $\frac{2}{4}, \frac{3}{4}\frac{5}{4}, \frac{6}{8}, \frac{7}{8}, \frac{9}{8}$, etc. In the suggested method, tempo induction will be examined in combination with the meter induction. In this direction, the number of appearances of the formed pair will be processed in terms of histogram and their ratio quality, in terms of *round of error*, will also be examined. The pair yielding the greater number of occurrences and the best *round of error* will be selected as the tempo and music meter pair.

The conducted experiments for 350 audio recordings revealed that, the algorithm retrieved successfully the tempo and music meter for the 95% of the cases.

In the context of pairwise periodicity induction, we advanced the proposed method and developed a variation based on a double histogram processing and a different ratio of correlated periodicities. The above application aimed at the development of a method that extracts two tempo periodicities from an audio recording that coincide with the pair of periodicities perceived by a human listener [7]. The proposed method took place in the MIREX international contest [8] and the results are exhibited in Table 2.

Tempo Deviation (%)	Both tempi correct (%)	At least one tempo correct (%)
2.5	54.95	89.60
5	59.9	92.57
8	60.89	94.06

Table 2. Performance of the algorithm for the Greek Traditional Dance.

5 Music summarization

In the context of exploring the periodicities of an audio recording, we propose a method for the discovery of the two most similar excerpts of an audio recording. We will consider these excerpts as the thumbnail of the audio recording and the suggested method explores another aspect of the SSM. The above excerpts are considered as the audio thumbnail of the recording.

In this sense, the *scale based MFCCs* is calculated and a dimensionality reduction takes place using the Singular Value Decomposition, (SVD) method. The SVD method is applied on the transpose, $\bar{c}^T$, of $\bar{c}$, i.e., $\bar{c}^T = USV$, where $U_{M\times L}$ and $V_{L\times L}$ are the projection matrices and $S_{L\times L}$ is the matrix of singular values and M is the number of short-term frames and L the number of output filters used. The first six rows of the transpose, U^T , of U, are finally selected as the feature sequence.

In the sequel, the *SSM* is generated from the first six rows of $\mathbf{U}^T$ and the Euclidean Distance. At a first step, the *SSM* is correlated with a rectangular window, w (size $D \times D$). The window has 1's on the main diagonal and zeros elsewhere. If (i, j) are the position indices of an element of *SSM*, the upper left corner of w is chosen to coincide with (i, j). The correlation result, $S(i, j)$, for $SSM(i, j)$ is therefore computed as follows:

$$S(i,j) = \sum_{d_1=0}^{D-1} \sum_{d_2=0}^{D-1} SSM(i+d_1, j+d_2) w(d_1, d_2) = \sum_{d=0}^{D-1} SSM(i+d, j+d).$$

At a second step, let $S(k, m)$ be the lowest value of S yielding the two segments representing the audio thumbnail. $S(k, m)$ resides on the diagonal with index $k - m$ and elements $\{S(k, m), S(k + 1, m + 1), \ldots, S(k + D - 1, m + D - 1)\}$ form a segment on the diagonal that defines the desired thumbnail. Parameter D controls the size of the thumbnail and is user defined, depending on the corpus under study.

6 Music retrieval by rhythmic similarity

In an attempt to retrieve recordings of rhythmic similarity, we propose a method based on SSM and parts of it's mean diagonals that we refer to with the term *rhythmic signature*. The similarity measurement between signatures of audio recordings is performed by means of a standard Dynamic Time Warping technique [9].

At a first step, the music signal is short-term processed to extract a sequence of *scaled based MFCCs*, i.e. $\mathbf{C} = [\underline{c}(1)\ \underline{c}(2)\ \ldots\ \underline{c}(N)]$, be the new sequence of MFCCs. In the sequel, $\mathbf{C}$ is long-term segmented with a moving long-term window (window length is 4 *secs* and step is 1 *sec*). To simplify notation, let $\mathbf{C_t} = [\underline{c_t}(1)\ \underline{c_t}(2)\ \ldots\ \underline{c_t}(M)]$, be the subsequence that corresponds to the t-th long-term window, where M is the window length measured in number of frames. The *SSM* is then calculated for each long-term window, using the Euclidean Distance metric. For the t-th long-term window, the mean value, $R_t(k)$, of each diagonal in the lower SSM triangle is computed, i.e., $R_t(k) = \frac{1}{M-k} \sum_{l=k}^{M} \left\| \underline{c_t}(l), \underline{c_t}(l-k) \right\|$, where k is the diagonal index and $\|.\|$ is the Euclidean distance function. Each R_t is treated as a signal. At a next step, the mean signal, R_μ, of all R_t's is computed, i.e., $R_\mu(k) = \frac{1}{T} \sum_{t=1}^{T} R_t(k)$, and then normalized to unity, where T is the number of long-term windows.

In what follows, we will refer to R_μ as the *rhythmic signature* of the music recording. The main idea behind this approach, is that, recordings with similar rhythmic characteristics are expected to yield "similar" signatures. Therefore, the next challenge is to devise a similarity measure for signatures.

If L is the number of music recordings in a corpus, L rhythmic signatures are first extracted. In order to measure similarity between signatures, a standard *Dynamic Time Warping* cost has been employed. As is the case with DTW techniques [3], a set of local path constraints needs to be first defined. In our study

we experimented with two types of constraints, i.e., *Sakoe-Chiba* and *Itakura* and adopted the former.

Precision %	Class 1	Class 2	Class 3	Class 4	Recall %	Class 1	Class 2	Class 3	Class 4
Class 1	**94.3**	3.2	1.7	0	Class 1	**94.3**	3.8	1.9	0
Class 2	3.8	**96.8**	0	0	Class 2	3.2	**96.8**	0	0
Class 3	1.9	0	**96.6**	10.9	Class 3	1.6	0	**90.3**	8.1
Class 4	0	0	1.7	**89.1**	Class 4	0	0	2.4	**97.6**

Table 3. Precision and recall for Greek Traditional corpus.

If a rhythmic signature is drawn from the corpus, its matching cost against the remaining L-1 signatures is calculated using the adopted DTW technique. This procedure yields $L-1$ cost values which are sorted in ascending order, with the lowest values indicating highest similarity.

Table 3 presents the confusion matrix for the four classes of Greek genres, where the leave-one-out method was applied. Table 3 reveals that, when only the lowest matching cost was examined, limited confusion occurred between the classes 3 and 4 and classes 1 and 2. Further experimentation revealed that, when the two lowest matching costs were taken into account, the confusion matrix remained the same within statistical confidence.

7 Modeling and tracking of rhythmic patterns.

In this work, we use *HMM*s to locate *rhythmic patterns* in music recordings by employing an *enhanced Viterbi* algorithm. The proposed method operates on the assumption that the music meter and a rough estimate of the tempo are known. It is assumed that tempo remains approximately constant throughout the recordings. To our knowledge, this is the first time that the problem of beat and meter tracking is addressed in the context of complex meters *without* any prior knowledge of any pattern location. Our focus is on complex meters, such as $\frac{7}{8}$, $\frac{9}{8}$, which appear in eastern folk music but also patterns of $\frac{2}{4}$, $\frac{3}{4}$ that are also frequently encountered in traditional dances. .

In our approach, a *rhythmic pattern* is modeled by means of a Hidden Markov Model, where each event of the pattern corresponds to a *HMM* state. In order to locate occurrences of such a pattern in a recording, the *HMM* is fed with overlapping segments of the feature sequence that has been extracted from the audio data and at a next step the extracted patterns (if any) are connected using a dynamic programming technique, thus creating a chain of *rhythmic patterns*. Finally, if are any patterns missing an additional HMM tracking step takes place using the previously extracted patterns a "seeds".

7.1 Modeling of Rhythmic patterns by means of HMM

Rhythmic structures can be considered to build upon fundamental *rhythmic patterns*. For example, recordings of music meter $\frac{7}{8}$ with *tempo* ranging from

200 - 290bpm, as is the case with a number of traditional music genres, are perceived as a *rhythmic pattern* of a sequence of [*dotted quarter note - quarter note - quarter note*]. Table 1 exhibits the patterns modeled in this study. This is also consistent with the performance of the accompaniment instruments and the singing voice in such recordings. To construct the corresponding *HMM*, each component of the above *rhythmic pattern* will be represented by a *HMM* state. Each state models by means of a *Gaussian pdf* with mean value, μ_i, the time duration of the respective event (within an allowable *tempo* fluctuation). As shown in Figure 1, for the example of $\frac{7}{8}$ we have three states each tuned to the respective event duration.

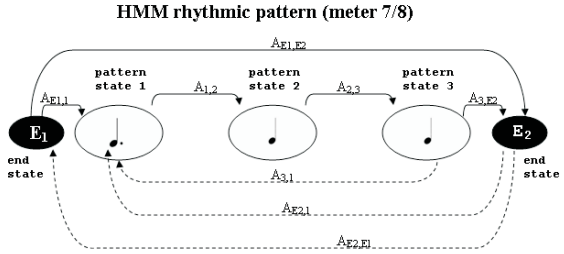

Fig. 1. 3-state HMM modeling a $\frac{7}{8}$ rhythmic pattern.

7.2 End States and Enhanced Viterbi Algorithm

In Figure 1 except from the three rhythmic *pattern* states, two more states are added, namely E_1, and E_2 (displayed in black). These states will be referred to as *end* states and are allowed to emit all the detected *IOI*s with a uniform probability. The physical meaning of these states is that the *HMM* can bounce between them whenever a sequence of *IOI*s does not conform with the pattern being modeled. On the other hand, the states that model the *rhythmic pattern* are assumed to emit *IOI*s following a Gaussian probability.

In *HMM* terminology [10], let $\lambda = \{\pi, \mathrm{A}, \mathrm{B}\}$, be the parameters of the *HMM* that models a rhythmic pattern. π_i is the initial state probability, $A_{S\times S}$ the state transition matrix, B_i the Gaussian probability distribution (*pdf*) of each *pattern* state, and S the number of states (including *pattern* and *end* states). Each Gaussian *pdf*, is associated with a *pattern* component (i.e., *dotted quarter*), with mean time duration μ_i and standard deviation σ_i,where time is measured in frames. The initial probabilities were set to $\pi_{=}[\frac{1}{2}\ \frac{1}{2}\ 0 \ldots 0]$, ($S$-1 zeros), forcing all paths to start from the first *end* state or the first *pattern* state. Furthermore, all self transition probabilities are set to zero, i.e., $A_{i,i} = 0$. The only allowable *right* to *left* transitions are those from the second *end* state and the last *pattern* state to the first *end* state and the first *pattern* state, marked with dashed arrows in Figure 1. This allows for tracking repetition in terms of *rhythmic patterns* if the long-term window is long enough [7].

As it is known [10], the standard *Viterbi* algorithm employs a *Type B* cost function for the generation of the *trellis* diagram. A *Type B* cost function takes into consideration *both* the transition costs between nodes $[i, j]$ ($A_{i,j}$ $B_j(t)$), as well as the accumulated node costs ($a_{t-1}(i)$). In our approach, a *Type T* cost function was used instead, that only accounts for the transition cost between nodes. It is worth mentioning that a *Type T* cost retains the *Markovian* nature of the *trellis* diagram [11]. In *Markov* model terminology the delta variable [10] reduces to:

$$\delta_t(i, j) = A_{i,j} B_j(t) \tag{2}$$

By eliminating the forward probability from Eq. 2, this cost function takes into account only the "local" activity of the most recent transition. If the *HMM* enters several times the *end* states before entering the *pattern* states, this will not affect local high probability transitions between *pattern* states which indicate that the pattern has been found.

To find the best state sequence, $Q = \{q_1, q_2, \ldots q_T\}$ for each long-term segment the arguments that maximize the forward variable equation are first stored in a two dimensional array ψ, as $\psi(j, t)$

$$\psi(j, t) = argmax\ [\delta_t(i, j)],\ 1 \leq i \leq S,\ i \neq j \tag{3}$$

At a next step, a backtracking procedure is applied on every node that corresponds to the last state of the *rhythmic pattern*, irrespective of time instance. This is expected to yield a number of paths. In order to select the best one (with the highest probability), the path probabilities have to be computed. To this end, if $Q = \{q_1, q_2, \ldots q_T\}$ is an extracted path, the associated probability is calculated from the equation:

$$p_{model} = \prod_{\forall q \in Q} a_t(q),\ \textit{and } q \textit{ not an end state.} \tag{4}$$

As shown in the above Equation (4), the *end* states do not participate in the calculation of the pattern recognition probability since they do not belong in the *rhythmic patterns* modeled by the *HMM*s.

Due to the nature of polyphonic music, it is obvious that the onsets returned during the feature extraction process will outnumber the onsets corresponding to the *correct* beat locations. To address the above problem, an *enhancement* of the *Viterbi* algorithm was employed. Let us consider the onset sequence F for an audio region, i.e.:

$$\begin{aligned} F = \{\ldots, a(m-3), O_{z_{m-3}}, a(m-2), O_{z_{m-2}}, \\ a(m-1), O_{z_{m-1}}, a(m), O_{z_m}, a(m+1), O_{z_{m+1}}, \ldots\} \end{aligned}$$

Let $a(m)$ and $a(m-3)$ be two *correct* onsets with two *false* ones, $[a(m-2)$, $a(m-1)]$ in between. Their corresponding durations of $[a(m-2), a(m-1)]$ are $[O_{z_{m-2}}, O_{z_{m-1}}]$. Although $a(m-3)$ is a *correct* onset, it's corresponding duration $O_{z_{m-3}}$ is erroneous, due to the presence of the events $a(m-2), a(m-1)$.

Taking into account the zero components, the correct duration can be derived as $\sum_{i=1}^{3} O_{z_{m-i}}$. In this way, we offer to the *HMM* the possibility to eliminate *false* onsets and keep the *correct* ones, while searching for the optimal path and if a lower cost (higher probability) is achieved by eliminating events, the *Viterbi* is given the means to do it. In other words, the cost now becomes "context" dependent. This context dependency of the *Viterbi* algorithm leads to the modification of Equation (2) as:

$$\hat{\delta}_t(i,n,j) = A_{i,j}\hat{B}_j, \tag{5}$$

where: $\hat{B}_j = B_j(\sum_{d=t-n+1}^{t} O_{z_d})$, where n is the index of the zero component being added and D the maximum number of observations allowed to be summed. The maximum number of observation symbols over which a state is allowed to sum, depends upon the tolerance of each state mean duration variation $\Delta\mu_i$. This is expressed as: $\forall i \in$ *pattern* states, $\sum_{d=1}^{D} O_{z_d} \leq \Delta\mu_i$. In this work, a constant state duration variation $\Delta\mu_i \simeq 20\%\mu_i$ was allowed, based on extensive experimentation. The forard variable Equation is now transformed to:

$$\hat{a}_t(j) = \max_{1\leq t\leq T, 1\leq n\leq D, 1\leq i\leq S, i\neq j} \left[\hat{\delta}_t(i,n,j)\right] \tag{6}$$

Unlike the *pattern* states, the *end* states are not allowed to sum consecutive onsets. This is justified by the fact that *end* states are not actually a part of the examined *rhythmic pattern*, but rather serve as "collectors" for erroneous and "off-beat" onsets. After the whole feature sequence has passed through the *HMM* the resulting pattern's locations are examined. Among the correct locations returned by the algorithm, false pattern locations may appear or missing ones. Toward this direction, a dynamic programming technique is developed and a cost grid is formed in order to select those patterns that lie in succession, thus forming a "winning" chain pattern. Furthermore, if missing patterns exist the patterns in the "winning" chain pattern serve as "seeds" for the additional extraction of patterns. The results for the modeled patters are exhibited in Table 4 for the mean tempo case and the intentionally introduced erroneous tempo case.

Table 4. Evaluation results on a stage basis using the proposed method.

8 Conclusions

This thesis presented several content-based analysis methods for the extraction and discovery of rhythmic characteristics from audio recordings. Our study has focused on non-western corpora that exhibit a wide variety of rhythms and rhythmic patterns. The developed techniques exhibit satisfactory results for the

studied corpora. The future work will mainly focus on the automatic modeling of the rhythmic patterns directly from the audio recording and the automatic extraction of the audio thumbnail size.

References

1. G. E. Metallinos, "Percussions and Greek Traditional Music Rhythms", Polyrhythmia, Athens, 1993.
2. S. Karas, "Methods for Greek Music, Theoritikon", Athens, 1982.
3. S. Theodoridis and K. Koutroumbas, "Pattern recognition", Academic Press, 3d Edition, 2006.
4. Slaney, M., "The Auditory Toolbox", Apple Technical Report #45, 1993, Apple Computer Inc, 1 Infinite Loop, Cupertino, CA 95014.
5. Aggelos Pikrakis, Iasonas Antonopoulos, and Sergios Theodoridis, "Music Meter and Tempo Tracking from Raw Polyphonic Audio", in ISMIR Proceedings, 2004, Barcelona, Spain, October 2004.
6. J. Foote, "Visualizing Music and Audio using Self- Similarity", in Proceedings of ACM Multimedia 99, pp. 77-80 Orlando, FL, USA, ACM Press, 1999.
7. Iasonas Antonopoulos, Aggelos Pikrakis and Sergios Theodoridis, "Self-Similarity Analysis Applied on Tempo Induction from Music Recordings", in Journal of New Music Research, Volume 36, Issue 1 March, 2007, pages 27-38.
8. MIREX, 2006, URL: *http://www.music ir.org/mirex2006/index.php* .
9. Iasonas Antonopoulos, Aggelos Pikrakis, Sergios Theodoridis, Olmo Cornelis, Dirk Moelants, Marc Leman, " Music Retrieval by Rhythmic Similarity applied on Greek and African Traditional Music", Proceedings of the 2007 International Conference on Music Information Retrieval and Related Activities - ISMIR 2007, September 23-27, 2007, Vienna, Austria.
10. L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in Proceedings of the IEEE, Vol. 77, No. 2, 1989.
11. John G.Proakis John R. Deller Jr and John H.L.Hansen, "Discrete-time processing of speech signals", Prentice Hall, US Edition, 1987.