A NONLINEAR DYNAMIC MODELLING FOR SPEECH RECOGNITION USING RECURRENCE PLOT - A DYNAMIC BAYESIAN APPROACH

Satish Prabu Chandrasekaran
Digibee Microsystems,
DSP System,
Chennai - 600 004, India
satishprabu@ieee.org

ABSTRACT

The paper describes about a novel nonlinear feature extraction technique based upon Recurrence Plot (RP). This plot not only helps in visualizing the system dynamics but also can be quantified. The Recurrence Quantification Analysis (RQA) characterizes various aspects of a dynamic system and makes it a suitable technique for feature extraction. We have taken three prime quantification techniques namely Recurrence Rate, Entropy and Average Diagonal Length. The information about the system gets distributed in these quantities. Hence we need a model that is capable of taking into account the information from all the three RQA techniques. Dynamic Bayesian Networks (DBNs) can model these information very efficiently. For this purpose we have used Factorial Hidden Markov Model (FHMM) which is a special case of DBNs. The proposed method works well even in presence of noise when compared with the conventional technique.

Index Terms— Speech recognition, Dynamics, Bayes procedures, Hidden Markov models, Learning systems

1. INTRODUCTION

The structure of an utterance is determined by the combined effect of excitation source and vocal tract characteristics. Popular linear classical speech model based speech recognition methods uses the formant structure to quantify an utterance. But it is shown in [1] that given set of formant frequencies corresponding to a vocal tract configuration, other than this configurations there exists more number of different configurations resonating at those frequencies. The existence of multiple solutions implies the inherent nonlinearity in the speech production model.

The velocity of the air flow inside the larynx is unevenly distributed, with higher velocity near the walls of vocal tract and decreases towards the center[2]. This pragmatic situation overrides the plane wave propagation assumption of the classical physical model. Also there exists a nonlinear coupling of energy between different parts of the vocal tract. Clearly, the linear model overlooks or fails to model some of the natural phenomenon which really happens.

Also it is being shown in [3] that, when articulatory model takes into account the dynamics associated with convective flow and the viscous drag, it models the two dimensional wave propagation instead of planar propagation, can depict the speech production phenomenon in more faithful manner.

In addition, the display of chaotic behavior of the speech production mechanism bolsters the existence of the nonlinearity [4] in speech production. Also, the gaussian excitation assumption for non-voiced sounds need not hold true for all the cases. The above evidences show that more credible modelling of the speech could be done when the associated nonlinearities are taken into account. In this paper this is done with the aid of Recurrence Plot (RP) and Dynamic Bayesian Networks (DBN).

![Figure 1](image)

Figure 1. General block diagram of the proposed RQA based speech recognition system. RR - Recurrence Rate, ENTR - Entropy and L - Average diagonal length, see section 3.2. The DBN used here is a three layer Factorial HMM, see section 4

The organization of the paper is as follows, section 2 explains about modelling speech from a nonlinear dynamical perspective and representation of it in a phase space. Section 3 explains about Recurrence Plot & its quantification namely Recurrence Quantification Analysis (RQA) and also motivation for using it in speech analysis and recognition. Section 4 describes how to model the acoustic variation characterized by RQA and significance of DBN in modelling these variation. Section 5 gives the comparative results of the present method with conventional MFCC+HMM method. The results shows that at higher and lower SNR as well, performance of the proposed method is superior to that of the conventional MFCC+HMM method. Figure 1 shows the flow involved in proposed method.
2. MODELLING SPEECH PRODUCTION AS A NON-LINEAR DYNAMIC SYSTEM

It has been shown in [4] that nonlinearities found in speech could be analyzed from a nonlinear dynamic perspective. Hence in this paper we regard speech production mechanism as a dynamic system rather than a stochastic system.

Generally in case of a dynamical system the present output depends entirely on the previous set of values. That is, the output of the system at time \( t \) can be expressed as a function of previous outputs i.e.,

\[
y_t = f(y_{t-1}, y_{t-2}, \ldots)
\]  

(1)

In other words, there involves no element of chance and more formally, a dynamical system is characterized by its phase space and its time evolution law. The phase space represents all possible states of the system and the time evolution law determines the state of the system at a particular time \( t \).

In this paper we deal with speech which is a one dimensional signal. The dynamics associated with the system which produced the speech cannot be described properly by the observations made in time domain due to lack of dimensionality. Hence the phase space associated with that dynamical system has to be reconstructed and the process involved is termed as embedding. This process is based up on Time-Delay Embedding Theorem [5].

**Theorem 1** Given a dynamical system with a \( N \) – dimensional solution space and an evolving solution \( h(t) \), let \( x \) be some observation \( x(h(t)) \). Let us also define the lag vector (with dimension \( d \) and common time lag \( \tau \ ))

\[
x(t) = \{x_{t_0}, x_{t_1}, x_{t_2}, x_{t_3}, \ldots, x_{t_{N-1}}\} \tag{2}
\]

Then, under very general conditions, the space of vectors \( x(t) \) generated by the dynamics contains all of the information of the space of solution vectors \( h(t) \). The mapping between them is smooth and invertible. This property is referred to as diffeomorphism and this kind of mapping is referred to as an embedding. Thus, the study of the time series \( x(t) \) is also the study of the solutions of the underlying dynamical system \( h(t) \) via a particular coordinate system given by the observable \( x \).

3. RECURRENCE PLOTS (RP) FOR SPEECH FEATURE EXTRACTION

In this paper, we propose a method based on Recurrence Plots to characterize an utterance. A detailed discussion on the properties and application of the RP to a general nonlinear dynamical systems is given in [6]. The motivation for using RP for speech feature extraction are as follows

- A dynamical system can be well depicted by its state recurrence in the phase space. Hence by analyzing the state recurrent structure of the speech rendered by RP, can characterize the vocal tract.
- When dealing with speech time series for recognition, the length of the frame is supposed to be not greater than the phoneme’s duration. The RP analysis can be made even on short and non-stationary time series.

### Table 1. Expresses the relation between the state of the system and its corresponding recurrence plot texture

<table>
<thead>
<tr>
<th>System</th>
<th>RP texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stationary Systems</td>
<td>Homogeneous</td>
</tr>
<tr>
<td>Periodic</td>
<td>Diagonal Lines</td>
</tr>
<tr>
<td>Quasi-Periodic</td>
<td>Checkerboard Structure</td>
</tr>
<tr>
<td>Non-Stationary</td>
<td>Dulls away from LOI</td>
</tr>
<tr>
<td>Systems</td>
<td></td>
</tr>
<tr>
<td>Abrupt changes in System's Dynamics</td>
<td>White areas or bands</td>
</tr>
<tr>
<td>States strongly fluctuating</td>
<td>Isolated recurrence points</td>
</tr>
<tr>
<td>State varies slowly</td>
<td>Vertical and Horizontal lines</td>
</tr>
<tr>
<td>Bowed lines</td>
<td>System dynamics are changing</td>
</tr>
</tbody>
</table>

- There involves no loss of information when represented using the recurrence plot, since reconstruction of the time series is possible from the plot alone.
- The additional advantage of RPs are that the visual recurrent structures characterizing a system, can be quantified. So we can use them as the *acoustic feature vectors* of an utterance.

3.1. Recurrence Plot

Recurrence Plots (RPs) dates back to 1987 used by Eckmann et al. to visualize the behavior of trajectories of dynamical systems in phase space [7]. In this they have written: "... recurrence plots are rather easily obtained aids for the diagnosis of dynamical systems. They display important and easily interpretable information about time scales which are otherwise rather inaccessible."

Recurrence plot measures the recurrences of the state of the dynamical system and formally defined by a matrix

\[
R_{i,j}(\epsilon) = \Theta(\epsilon - ||x_i - x_j||) \quad i, j = 1, 2, 3, ..., N
\]

\[
R_{i,j} \equiv 1 \Leftrightarrow x_i \approx x_j \tag{3}
\]

where \( \Theta(\cdot) \) is Heaviside function, \( \epsilon \) a threshold distance, \( x_i \) is the \( i^{th} \) point in phase space. By definition \( R_{i,i} \equiv 1 \) \( \forall i \leq N \). The main diagonal is always one and it is called the **Line of Identity** (LOI).

The most important property of the RPs are that the texture found in the plot can reveal information about the system.

3.2. Recurrence Quantification Analysis (RQA)

- **Recurrence Rate (RR)**: This measure quantifies number of times the state re-occurred in the course, which can be given by the percentage of black dots. It can also be considered as a *generalized auto-correlation function*, as it describes higher-order correlations between the points of the trajectory. The significance of autocorrelation is well studied in speech processing literature. It can be calculated by the following expression
Figure 2. Recurrence Plot of voiced and unvoiced speech. With $\epsilon = 0.8$ and using Euclidean norm, dimension $N$ and delay $\tau$ both taken as 3. Refer [6]. Zoomed up for clarity

\[
RR(\epsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}(\epsilon)
\]

where $N$ is the number of speech samples, $R(\epsilon)$ is the recurrence matrix corresponding to a threshold of $\epsilon$

- **Entropy:** Shorter diagonals characterize the uncorrelated process while longer ones characterize deterministic process. Entropy determines the probability of finding a diagonal of specified length in the given recurrence plot.

\[
ENTR = - \sum_{l=l_{\min}}^{N} P(l) \ln(P(l))
\]

where $P(l)$ probability of finding a diagonal of length $l$ which can be calculated as $p(l)/K$ where $p(l)$ is the histogram of the diagonal length and $K$ gives total number of diagonals ($K = \sum_{l=l_{\min}}^{N} p(l)$)

- **Average diagonal length:** A diagonal line of length $l$ indicates that trajectories at some different time instants, but for same time steps, are closer. Also Eckmann et al. have stated that “the length of the diagonal lines is related to the largest positive Lyapunov exponent(LE)” if there is one in the considered system [7]. The LE quantifies the rate of convergence or divergence with positive LE denoting convergence and negative LEs denoting divergence. LEs are even used as features in nonlinear speech analysis [4].

\[
L = \frac{\sum_{l=l_{\min}}^{N} l p(l)}{\sum_{l=l_{\min}}^{N} p(l)}
\]

Figure 3. RQA of Vowel /aa/. The RQA with embedding delay $\tau = 3$ and dimension $N = 3$. Euclidean norm with $\epsilon = 0.8$, window size=40 and window length $w = 2$ and $l_{\min} = 2$. Refer [6]

4. ACOUSTIC MODELLING OF SPEECH USING DBNS

DBNs are generalization of the Bayesian Network Theory to dynamic process. The following primary distinction between HMMs and DBNs make DBNs a suitable candidate for modelling the speech, based on the RQA based feature extraction process. In a HMM the hidden state of the system is represented in terms of single discrete random variable that can take $K$ possible values. But in the case of DBNs the hidden state can be represented in terms of $M$ random variables, each of which can be discrete or continuous.

In the present method, the system is characterized by means of three quantities namely Recurrence Rate, Entropy, and Average Diagonal Length, which carry information individually. Another words the information about the state of the system gets distributed among the three. So similarly, the state of the system at each time instance should be decided by a collection of state variables corresponding to these quantities. **Factorial Hidden Markov Model (FHMM)** is one such model in which state of the system gets disseminated and the outcome at any time $t$ is decided based on all the disseminated states at that time. This FHMM is a specific case of DBNs.

4.1. Combining information using FHMM

In FHMM the state of the system $S_t$ at a time $t$ is represented as collection of $n$ state variables represented as

\[
S_t = [S_t^1, S_t^2, S_t^3 \ldots S_t^n]^T
\]

The combination of information can be done in two ways. The first method [8] is by the assumption that the observation $O_t$ at time $t$ is distributed as Gaussian with same covariance $C$ as that of the distributed states and the mean $\mu$ a weighted sum of the mean of the individual state variables at that time $t$. 

\[ p(O_t|S_t) \propto \exp\left\{ -\frac{1}{2}(O_t - \sum_{k=1}^{n} \mu^{(S^m)}_{t,k})^T C^{-1}(O_t - \sum_{k=1}^{n} \mu^{(S^m)}_{t,k}) \right\} \]

(8)

where \( \mu^{(S^m)} \) is mean corresponding to the state at time \( t \) and layer \( m \) and \( C \) denotes the common covariance matrix.

In the second method to combine the \( n \) individual information, we assume that the probability distribution of the observation \( O_t \) at time \( t \) is the product of the individual distributions of each layer. This technique is called "streamed" method [9].

\[ p(O_t|S_t) \propto -\frac{1}{2} \prod_{k=1}^{n} \exp\left\{ (M_m O_t - \mu^{(S^m)}_{t,k})^T C^{-1}(M_m O_t - \mu^{(S^m)}_{t,k}) \right\} \]

(9)

where \( M_m \) is the matrix which splits the vector into streams.

Generally DBNs can be trained using Expectation Maximization (EM) Algorithm.

5. SPEECH RECOGNITION EXPERIMENT RESULTS

The effectiveness of the proposed method was validated over TIMIT corpus. The training set contains 2000 utterance of mixed dialect and the testing set consists of 103 utterances. First the corpus was evaluated over MFCC+HMM based speech recognition system then evaluated on RQA+DBN system proposed in this paper.

The RQA based feature vectors represents the speech signal in a \( N \)-dimensional space which is very large. Hence Feature Selection process was carried out [10]. These acoustic features were trained / validated over FHMM. Baye’s Net toolbox (http://bnt.sourceforge.net) was used to build the FHMM.

The typical values used for the RQA feature extraction are delay of \( \tau = 3 \), dimension \( N = 5 \), Euclidean Norm was used, threshold \( \epsilon = 0.8 \), window size \( = 40 \), window length \( = 2 \), and \( \ell_{m,hr} = 3 \). In order to check its robustness, four types of noise (Gaussian, babble, car and airplane) were added in different proportions and validated. The significance of the proposed method is evident form Figure 4

6. CONCLUSION AND FUTURE WORKS

The proposed method shows comparatively better performance at low SNR. Still other quantification techniques and the associated complexity of the RF based system are yet to be explored also benchmarking it with other multi-feature speech recognition system and kept as future works.

7. REFERENCES


