Traffic Pattern Prediction Based Spectrum Sharing for Cognitive Radios

Xiukui Li and Seyed A. (Reza) Zekavat
Michigan Technological University, Houghton, MI 49931
USA
{xiuli,rezaz}@mtu.edu

1. Introduction

In this chapter, we introduce different traffic prediction techniques and discuss the process of evaluating channel availability through predicting traffic pattern of primary users for cognitive radios. When cognitive and non-cognitive users share the licensed spectrum, compared with secondary users (cognitive users), primary users have higher priority in using licensed channels. Therefore, whenever a primary user is detected, secondary users must vacate the relevant channels or decrease their transmitted power to reduce the interference on primary users. However, in some situations, due to the activities of primary users, secondary users may need to switch to other available channels, terminate communication, or reduce the transmitted power frequently. This leads to temporal connection loss of secondary users. In addition, if secondary users can not vacate a channel in a timely manner, it would interfere with primary users. To reduce the temporal connection loss and interference on primary users, secondary users should avoid using the channels which would be claimed by primary users with a high probability within a given time period. A solution to this problem is to enable secondary users to evaluate the channel availability through predicting the traffic pattern of primary users. This chapter discusses the methods of evaluating the probability of channel availability for secondary users, and examines the probability that there is a successful secondary user’s transmission at one time instance. In addition, the cooperative prediction is briefly introduced. Finally, the applications of traffic pattern prediction technique to spectrum sharing are discussed.

2. Traffic Model

Traffic data patterns can be classified as (Haykin, 2005): 1) deterministic patterns: for example, each primary user (e.g., TV transmitter) is assigned a fixed time slot for transmission, and when it is switched off, the frequency band is vacated; 2) stochastic patterns: for example, the arrival times of data packets are modeled as a Poisson process, while the service times are modeled as exponentially distributed, depending on whether the data are of packet-switched or circuit-switched. Note that, in common channel signaling network, the exponential distribution drastically underestimates the proportion of short calls that do not last longer than the mean holding time (Bolotin, 1994). In general, the traffic stochastic parameters vary slowly. Hence, they can be estimated using historical data. The traffic model built on historical data
enables secondary users to predict the future traffic pattern of primary users (Li & Zekavat, 2008). An overview of traffic modeling is provided in (Frost & Melamed, 1994). It discusses Markov modulated traffic models, autoregressive traffic models and self-similar traffic models, etc. In addition, the authors in (Adas, 1997) discuss different traffic models in telecommunication networks.

2.1 Traffic Model for Voice Communication
A speech process can be modeled as a two-state Markov chain which alternates between talk spurt and silent periods (Li, 1990); (Gruber, 1981). The authors in (Hong & Rappaport, 1986) propose a traffic model for cellular mobile radio telephone systems. The basic system model in (Hong & Rappaport, 1986) assumes that the new call origination rate is uniformly distributed over the mobile service area, and the channel holding time is approximated to a negative exponential distribution.

Considering a one-dimensional mobile system with cells in series (e.g., in highways), the authors in (Pavlidou, 1994) uses two-dimensional state diagrams to analyze the traffic in the mixed media cellular system. It assumes four Poisson arrival streams are entering each cell, which are originating new voice calls, originating new data packets, hand-off voice calls and hand-off data packets. The authors in (Leung et al., 1994) also consider the circumstance of a one-way, semi-infinite highway. With the assumption that there are an infinite number of channels available, they present a deterministic fluid model and a stochastic traffic model for a wireless network along the highway.

2.2 Traffic Model for Video Data
The authors in (Dawood & Ghanbari, 1999) provide a summary for traffic models of video data. (Lucantoni et al., 1994) proposes to model a single video source as a Markov renewal process whose states represent different bit rates. Some other models including Markov Modulated Fluid Flow (MMFF) model (Maglaris et al., 1988), Markov Modulated Poisson Process (MMPP) (Skelly, 1993), and AutoRegressive AR(1) stochastic model (Maglaris et al., 1988) are proposed to address the basic characteristics of the variable bit rate traffic. The MMFF and MMPP are suitable for queueing analysis of packet switched networks. The AR(1) stochastic model primarily characterizes the inter-frame source bit-rate variations and correlation.

For variable bit-rate traffic, the authors in (Knightly & Zhang, 1997) introduce a new deterministic traffic model called deterministic bounding interval-length dependent (D-BIND) to capture the multiplexing properties of bursty streams. For a large-scale satellite network simulation, (Ryu, 1999) proposes: 1) a discrete autoregressive process for MBone (“multicast backbone”) video source modeling; 2) the superposition of fractal renewal processes (Sup-FRP) model for Web request arrivals, and, 3) a generalized shot-noise-driven Poisson point process (GSNDP) for aggregate traffic flow modeling.

2.3 Traffic Model for General Packet Data
Based on an analysis of Internet protocols for data communication, (Anderlind & Zander, 1997) proposes a simple model for future data traffic in wireless radio networks. Model parameters are selected to describe traffic from the Worldwide Web (WWW) access and from distributed file systems. A multilayer Markov model is considered in (Filipiak, 1992) for arrivals of calls, bursts, and packets to fast packet switching system, where the multilayer refers to call layer, packet layer and burst layer.
3. Traffic Prediction Techniques

Different methods and models are proposed to forecast the traffic of different networks. Generally, network traffic has a mix of self-similarity, Short Range and Long Range Dependencies (SRD and LRD) (e.g., (Kleinrock, 1993); (Paxon & Floyd, 1995); (Leland et al., 1994); (Jiang et al., 2001)). There exist concentrated periods of low activity and high activity (i.e., burstiness) in the network traffic (Feng, 2006).

An accurate predictor needs to capture the traffic characteristics such as SRD and LRD, self-similarity and nonstationarity (Sadek & Khotanzad, 2004a). The prediction models can be categorized as stationary and non-stationary ones (Adas, 1997). The stationary models include traditional models such as Poisson, Markov and autoregressive (AR), and self-similar models such as fractional autoregressive integrated moving average (FARIMA). Nonstationary models include artificial neural network (ANN) such as fuzzy neural network (FNN).

These models can be applied to predict different types of traffic data in Ethernet, Internet, etc. (Sadek et al., 2003); (Shu et al., 1999); (Hall & Mars, 2000). In the following subsections, some prediction techniques are introduced including ARIMA based traffic forecasting, application of neural network in traffic forecasting, least mean square based traffic forecasting, etc.

3.1 ARIMA based Traffic Forecasting

Traditional models, such as autoregressive (AR) and autoregressive moving average (ARMA), can be used to predict the high-speed network traffic data (Adas, 1997); (Sang & Li, 2002), and they can capture the short range dependent (SRD) characteristic of traffic. For example, considering the traffic data from a regional emergency communication center, (Chen & Trajkovic, 2004) classifies network users into user clusters, and then predicts the network traffic using the SARIMA (seasonal autoregressive integrated moving-average) model. SARIMA is discussed in detail in (Brockwell & Davis, 2002). For a campus-wide wireless network (IEEE 802.11 infrastructure), (Papadopouli, 2005) uses the Yule-Walker method of ARIMA model to forecast the traffic at each wireless access point (AP) in an hourly timescale.

If a time series is non-stationary, but the first difference of the time series is stationary, ARIMA models can be used to characterize the dynamics of such a process. (Basu, 1996) indicates that appropriately differenced time-series generated from Internet traffic traces can be modeled as Auto-Regressive-Moving-Average (ARMA) processes. (Krithikaivasan et al., 2004) also proposes to use the ARIMA models to predict network traffic. Although ARIMA models can be used to model a class of non-stationary data, however, they cannot be applied to predict the network traffic which possesses the Long Range Dependent (LRD) characteristics.

In ARIMA models, the difference parameter \(d\) is restricted to integer values. In Fractional Autoregressive Integrated Moving Average models (FARIMA), this difference parameter can be a fraction. FARIMA model has the ability to capture both SRD and LRD characteristics of traffic (Corradi et al., 2001); (Shu et al., 1999). (Shu et al., 1999) also proposes using fractional ARIMA (FARIMA) to capture the self-similarity of network traffic. However, FARIMA model is time-consuming (Feng, 2006). (Sadek & Khotanzad, 2004a) discusses a two-stage predictor. It combines two different models, namely FARIMA and FNN. FARIMA captures the self-similarity, and FNN captures the non-stationarity. The combination of FARIMA and FNN enhances the prediction accuracy.

The authors in (Papagiannaki et al., 2005) model the evolution of IP backbone traffic at large time scales for long-term forecasting of Internet traffic. The aggregate demand between any two adjacent points is modeled as a multiple linear regression model. Short-term (e.g., sec-
onds or minutes) forecasting of Internet traffic is addressed in (Basu, 1999); (Sang, 2000); (Papadopouli et al., 2006).
The authors in (Randhawa & Hardy, 1998) model the VBR sources as AutoRegressive AR(1) Modulated Deterministic Arrival process which characterizes the inter-frame as well as the intra-frame bit rate variations, and they model the call arrival process as conventional birth-death Markov Process. The future traffic is predicted using a combination of linear prediction and transient state analysis of birth-death Markov Process.

3.2 Application of Neural Network in Traffic Forecasting
The authors in (Zhao et al., 2004) propose two neural network models for traffic forecasting in two situations. One addresses the forecasting of hourly traffic of the next day based on past observations, and the other focuses on the peak load prediction of the next day. Based on Time Series Forecasting (TSF), (Cortez et al., 2006) uses a Neural Network Ensemble (NNE) to predict the TCP/IP traffic.
Neural network based traffic prediction approach is complicated to implement. The accuracy and applicability of the neural network approach in traffic prediction is limited (Hall & Mars, 2000). Artificial Neural Network (ANN) can capture the non-linear nature of network traffic and the relationship between the output and input theoretically (Hansegawa et al., 2001); (Khotanzad & Sadek, 2003); (Lobejko, 1996), however, it might suffer from over-fitting (Doulamis et al., 2003).
The machine learning technique called support vector machine (SVM) can be applied to pattern recognition and other applications such as regression estimation. (Feng, 2006) employs the SVM to forecast the traffic in WLANs. It studies the issues of one-step-ahead prediction and multi-step-ahead prediction without any assumption on the statistical property of actual WLAN traffic.

3.3 Least Mean Square based Traffic Forecasting
Normalized Least Mean Square (NLMS) based prediction approaches are proposed for online variable bit-rate video traffic prediction. They do not require any prior knowledge of the video statistics (Adas, 1998). The authors in (Liu & Mao, 2005) propose a time-domain NLMS based prediction scheme, and a wavelet-domain NLMS based adaptive prediction scheme for video traffic prediction which exploits the redundant information in the wavelet transform coefficients for more accurate prediction (traffic is decomposed into wavelet coefficients and scaling coefficients at different timescales).

3.4 Other Traffic Forecasting Algorithms
The authors in (Kohandani et al., 2006) present a new forecasting technique called extended structural model (ESM) which is derived from the basic structural model (BSM). This technique replaces the constant parameters in BSM with variables and use the steepest descent search algorithm to find the values for these variables to minimize the mean absolute percentage error (MAPE).
Based on classical queueing theory, (Filipiak & Chemouil, 1987) derives discrete-time stochastic models and propose methods based on those models to forecast the number of congested trunk groups. The authors in (Liang, 2002) show that the ad hoc wireless network traffic is self-similar, and they apply a fuzzy logic system to ad hoc wireless network traffic forecasting. Fuzzy logic systems (FLSs) have been widely used in time-series forecasting (e.g., (Liang & Mendel, 2000)).
<table>
<thead>
<tr>
<th>Traffic prediction techniques</th>
<th>Advantage</th>
<th>Disadvantage</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>Can capture the short range dependent (SRD) characteristic; can model a class of non-stationary data.</td>
<td>Cannot capture the Long Range Dependent (LRD) characteristic; the difference parameter ( d ) is restricted to integer values.</td>
<td>WLAN, Internet traffic, etc.</td>
</tr>
<tr>
<td>FARIMA</td>
<td>Can capture both SRD and LRD characteristics; can capture the self-similarity; the difference parameter ( d ) can be a fraction.</td>
<td>Time-consuming</td>
<td>Ad hoc wireless networks; Ethernet, etc.</td>
</tr>
<tr>
<td>Neutral Network based</td>
<td>Can capture the non-stationarity; can capture the non-linear nature of network traffic.</td>
<td>Complicated to implement; the accuracy and applicability is limited.</td>
<td>Internet traffic (video traffic), etc.</td>
</tr>
<tr>
<td>Least Mean Square based</td>
<td>Do not require any prior knowledge of the traffic statistics.</td>
<td>Slow convergence with the time-domain NLMS based algorithms; High complexity with the wavelet-domain NLMS based algorithms.</td>
<td>Internet traffic (video traffic), etc.</td>
</tr>
</tbody>
</table>

Table 1. Comparison of traffic prediction techniques.

The authors in (Sadek & Khotanzad, 2004b) present a parameter estimation procedure for the k-factor GARMA model (a generalized form of FARIMA). They use an adaptive prediction scheme to enable the model to capture the non-stationary characteristic and deal with any possible growth in the future traffic. The results indicate that the performance of the k-factor GARMA model outperforms the traditional AR model. GARMA models can be used to model a time series with SRD and LRD characteristics, and they also can model the periodicity of a time series with fewer parameters than ARMA models. The authors in (Ramachandran & Bhethanabotla, 2000) use the GARMA framework to measure the periodicities of Ethernet traffic.

The advantages and disadvantages of some traffic prediction techniques are compared in Table 1.

4. Probability of Channel Availability

Currently, most of the spectrum is assigned to GSM and CDMA networks (Brodersen, 2004). For CDMA networks (e.g., CDMA 2000, UMTS), the code division access makes the coexistence of primary users and secondary users difficult unless FDMA-CDMA systems are used (Pezeshk & Zekavat, 2003). In that case, secondary users may use the available spectrum only if the entire CDMA channels are available. Here, it is assumed that secondary users can find the idle channels for communication which are licensed to GSM networks.
4.1 Channels in GSM Networks

GSM uses TDMA techniques to provide multiple access for mobile users (Rappaport, 2002). In GSM, the available forward and reverse frequency bands are divided into 200kHz wide channels called ARFCNs (Absolute Radio Frequency Band Numbers). ARFCN is time shared between 8 subscribers using TDMA (there are eight time slots per TDMA frame). The combination of a time slot number and an ARFCN constitutes a physical channel for both the forward and reverse link. Each physical channel in a GSM system can be mapped into different logical channels at different times. There are two types of GSM logical channels, called traffic channels (TCHs) and control channels (CCHs).

In GSM systems, there are control messages transmitted over control channels between subscribers and base stations even if no call is in progress. There are three main control channels: the broadcast channel (BCH), the common control channel (CCCH), and the dedicated control channel (DCCH). The BCH and CCCH forward control channels are allocated time slot 0 (TS0). In the same frame, the time slot 1, TS1, through 7, TS7, still can carry regular traffic. The DCCH may exist in any time slot and on any ARFCN except TS0 of the BCH ARFCN. However, there are three types of DCCHs in GSM: (i) the stand-alone dedicated control channels (SDCCHs) which can be considered as intermediate and temporary channels for accepting a call from the BCH and holding the traffic while waiting for the base stations to allocate a TCH channel; (ii) the Slow, and, (iii) the Fast Associated Control Channels (SACCHs and FACCHs) which are used for supervisory data transmission between mobiles and base stations during a call. In summary, for the three control channels, BCH and CCCH only use TS0 in each frame and DCCHs are call related. Therefore, if there is no call in progress, secondary user can find an available channel among TS1-TS7 to use.

To detect an unused physical channel, a secondary user needs to tune to a forward channel (ARFCN) and synchronize with the base station. For a TDMA frame, a secondary user can detect whether the time slots TS1-TS7 are used or not by analyzing the data bursts. Here, secondary users are assumed to have the capability to remove the effect of dummy burst which are used as filler information for unused time slots on the forward link.

4.2 Channel Availability Evaluation

Generally, multiple channels are licensed to primary users. Secondary users can select any idle channel for communication. However, once a secondary user detects the appearance of primary users, it should vacate the channel immediately to allow primary users to continue using that channel. Therefore, the activities of secondary users do not affect the traffic distribution of primary users. Primary users keep their normal activities regardless of the existence of secondary users. Here, the number of channels licensed to primary users is assumed to be constant. In addition, with a fair scheduling policy, primary users utilize all licensed channel equally.

Traffic pattern prediction enables secondary users to estimate the channel utilization in a near future. For voice communications, two crucial factors in the traffic pattern are call arrival rate and call holding time. To estimate the utilization of one channel, secondary users can predict or estimate the call arrival rate and call holding time of primary users that use this channel. Then, according to the prediction and/or estimation results, secondary users are able to evaluate the probability that the channel would be available for a given time period. By comparing the evaluated probability with some threshold, secondary users can decide whether to use this channel.
In general, the traffic of voice communication in traditional wireless networks is periodic with a specific period $T$, e.g., $T = 24$ hours (one day), and it follows similar pattern in each period (Shu et al., 2003). An example of the call arrival process is shown in Fig. 1.

The call arrivals of primary users can be considered to follow non-homogeneous Poisson process $\{A(t), t \geq 0\}$ (Snyder & Miller, 1991). The rate parameter for the process $\{A(t)\}$ is $\lambda(t)$. $\lambda(t)$ may change over time. The expected call arrival rate between the time $t_1$ and $t_2$ is:

$$\lambda_{t_1,t_2} = \int_{t_1}^{t_2} \lambda(t)dt. \quad (1)$$

Thus, the number of call arrivals within the time interval $(t, t + \tau]$ follows a Poisson distribution with the parameter $\lambda_{t,t+\tau}$, i.e.,

$$P\{A(t+\tau) - A(t) = k\} = \frac{e^{-\lambda_{t,t+\tau}}(\lambda_{t,t+\tau})^k}{k!}, k = 0, 1, \ldots \quad (2)$$

Evenly dividing one traffic period into 24 time intervals $(t_n, t_{n+1}]$ ($n = 0, 1, \ldots, 23$), then, the time duration $T_d$ ($T_d = t_{n+1} - t_n$) for one time interval is one hour. For call arrival rate, a common metric employed in the telecommunication industry is the hourly number of calls (Chen & Trajkovic, 2004). Thus, the rate parameter $\lambda(t)$ can be assumed to maintain constant value $\frac{\lambda_n}{T_d}$ in each time interval $(t_n, t_{n+1}]$, i.e.,

$$\lambda(t) = \frac{\lambda_n}{T_d}, t \in (t_n, t_{n+1}] \quad (3)$$

where, $\lambda_n$ is the total number of call arrivals in the time interval $(t_n, t_{n+1}]$.

Considering $t_n < t < t + \tau \leq t_{n+1}$, i.e., $t$ and $t + \tau$ are within the same time interval $(t_n, t_{n+1}]$, the expected call arrival rate within $\tau$ is:

$$\lambda_{t,t+\tau} = \frac{\lambda_n}{T_d} \tau. \quad (4)$$
Similarly, using (1) and (3), \( \lambda_{t,t+\tau} \) can be obtained when \( t \) and \( t + \tau \) are within different time intervals or within different periods.

When a secondary user finds an idle channel and intends to start transmission over this channel, first, it predicts the number of primary user call arrivals in the current time interval \((t_n, t_{n+1})\). Using (4), the secondary user obtains the call arrival rate of primary users, \( \lambda_{N_{ch}} \), within its oncoming call holding time \( \tau_{ch} \) (the call holding time of its next call). Then, the secondary user evaluates the probability, \( P_{na} \), that no primary user would occupy the channel within the call holding time \( \tau_{ch} \). According to (2), \( P_{na} \) corresponds to:

\[
P_{na} = e^{-\frac{\lambda_{N_{ch}} \tau_{ch}}{T_d}} = e^{-\frac{\lambda_{N_{ch}}}{T_d}}.
\]  

The oncoming call holding time \( \tau_{ch} \) for a secondary user is a random variable and it is hard to predict. \( \tau_{ch} \) in (5) can be replaced with the average call holding time \( \bar{T} \) of the secondary user. For a secondary user, \( \bar{T} \) can be calculated based on the cumulative total call holding time \( T_d \) and the total number of calls \( N_c \) within a time duration, i.e., \( \bar{T} = \frac{T_d}{N_c} \).

To evaluate the probability that a channel is available for a given time period, secondary users should consider three scenarios discussed in the following:

Case 1 and 2 (see Fig. 2): Primary users end transmission at time \( t_1 \), and a secondary user starts transmission at \( t_2 \), \( t_n < t_1 < t_2 \leq t_{n+1} \).

In these two cases, the secondary user only needs to evaluate the probability \( P_i \) that no primary user arrives within \( \bar{T} \), \( i = 1, 2 \). According to (5), \( P_i \) corresponds to:

\[
P_i = e^{-\frac{\lambda_{N_{ch}}}{T_d} \bar{T}}, i = 1, 2.
\]  

Case 3 (see Fig. 2): One primary user starts transmission over a channel at time \( t_0 \), and it ends transmission at \( t_2 \). A secondary user intends to start transmission at \( t_1 \) over this channel (no other channels are available), \( t_n < t_0 < t_1 < t_2 \leq t_{n+1} \), i.e., at time \( t_1 \), the primary user call is still in progress. In this case, the secondary user is assumed to be capable of suspending its call for some time duration \( T_{sp} \) (\( T_{sp} > 0 \)). Otherwise, the call of secondary users would drop (see (Li & Zekavat, 2009) for details).

In summary, to evaluate the probability that the channel would be available within \( \bar{T} \), secondary users should be able to predict the number of primary user call arrivals in the corresponding time interval. In addition, secondary users should be able to obtain the PDF of call holding time distribution of primary users (for Case 3).

### 4.3 Call Arrival Prediction Algorithms

In Section 4.2, the rate parameter \( \lambda(t) \) for the process \( \{A(t)\} \) is assumed to be constant, \( \frac{\lambda_n}{T_d} \), in the time interval \((t_n, t_{n+1}] \) (\( T_d = t_{n+1} - t_n, n = 0, 1, \cdots, 23 \)), and \( \lambda_n \) is the total number of call arrivals in the time interval \((t_n, t_{n+1}] \). Therefore, to estimate the call arrival rate of primary
users within a given time period, secondary users need to predict the number of call arrivals in the corresponding time interval. Denoting the time interval \((t_n, t_{n+1}]\) in the \((m + 1)^{th}\) period as \((t_n + mT, t_{n+1} + mT]\), and the corresponding number of call arrivals in this time interval as \(\lambda_{n+mT}\), the set of observations of the number of call arrivals in different time intervals of different periods can be considered as a discrete-time series \(\{\lambda_t\} (t = 0, 1, \cdots)\). Thus, secondary users can predict the call arrivals of primary users by using the known observations of the number of primary user call arrivals in the past. Here, we use SARIMA model as an example to discuss one-step prediction of the number of call arrivals in a time interval, i.e., predicting the number of call arrivals \(\lambda_{t+1}\) given the known observations of call arrivals \(\lambda_i, i \in \{1 \cdots l\}\). \{\lambda_t\} is a seasonal ARIMA \((p, d, q) \times (P, D, Q)_s\) process with period \(s\) if the differenced series \(Y_t = (1 - B)^d(1 - B^s)^D \lambda_t\) is a causal ARMA process defined by:

\[
(1 - \phi_1 B - \cdots - \phi_P B^P) \left(1 - \Phi_1 B^s - \cdots - \Phi_P (B^s)^P\right) Y_t = (1 + \theta_1 B + \cdots + \theta_q B^q) \left(1 + \Theta_1 B^s + \cdots + \Theta_Q (B^s)^Q\right) Z_t,
\]

where, \(p\) and \(P\) are the non-seasonal and seasonal autoregressive orders, respectively; \(q\) and \(Q\) are the non-seasonal and seasonal moving average orders, respectively, and \(d\) and \(D\) are the numbers of the regular and seasonal differences required. In addition, \(\phi_1, \cdots, \phi_p, \Phi_1, \cdots, \Phi_P, \theta_1, \cdots, \theta_q, \Theta_1, \cdots, \Theta_Q\) are coefficient parameters. \(B\) is the backward operation, i.e., \(B \lambda_t = \lambda_{t-1}\), and \(Z_t \sim N(0, \sigma^2)\) (Brockwell & Davis, 2002).

With the simulated field data of \(\{\lambda_t\}\), the autocorrelation of \(\{Y_t\}\) with different \(d\) and \(D\) is calculated and it is found that \(d = 1\) and \(D = 1\) make the process \(\{Y_t\}\) stationary. With period \(s = 24\), the differenced observation \(Y_t\) corresponds to:

\[
Y_t = (1 - B)(1 - B^{24}) \lambda_t = \lambda_t - \lambda_{t-1} - \lambda_{t-24} + \lambda_{t-25}.
\]

Rearranging (8),

\[
\lambda_t = Y_t + \lambda_{t-1} + \lambda_{t-24} - \lambda_{t-25}.
\]

Considering \(h\)-step prediction, and setting \(t = l + h\):

\[
\lambda_{l+h} = Y_{l+h} + \lambda_{l+h-1} + \lambda_{l+h-24} - \lambda_{l+h-25}.
\]

Using \(P_l\) to denote the best linear predictor, according to (10),

\[
P_l \lambda_{l+h} = P_l Y_{l+h} + P_l \lambda_{l+h-1} + P_l \lambda_{l+h-24} - P_l \lambda_{l+h-25}.
\]

After calculating \(P_l \lambda_{l+h}\) of \(Y_{l+h}\), the prediction \(P_l \lambda_{l+h}\) of \(\lambda_{l+h}\) can be computed recursively by noting that \(P_l \lambda_{l+1-j} = \lambda_{l+1-j}\) for \(j \geq 1\). For one-step prediction \((h = 1)\), according to (11), \(P_l \lambda_{l+1}\) corresponds to:

\[
P_l \lambda_{l+1} = P_l Y_{l+1} + \lambda_l + \lambda_{l-23} - \lambda_{l-24}.
\]

The linear prediction of ARMA process \(\{Y_t\}\) \((P_l Y_{l+h})\) can be implemented by the innovations algorithm (Brockwell & Davis, 2002).
Some algorithms discussed in Section 3 can also be used to implement the call arrival prediction. For other traffic types in various applications such as image and video, different algorithms can be selected to forecast the future traffic.

If the current time is within the time interval \((t_{n+1}, t_{n+2}]\), given the predicted result \(\hat{\lambda}_{n+1}\), for Case 1 and 2 discussed in Section 4.2, the probability \(P_i\) that a channel would be available to the secondary user corresponds to:

\[
P_i = e^{-\frac{\hat{\lambda}_{n+1}}{T_i}}, i = 1, 2,
\]

where, \(\hat{\lambda}_{n+1}\) is the predicted result of \(\lambda_{n+1}\), and \(T\) is the average call holding time of the secondary user.

### 4.4 Decision of Channel Availability

Introducing traffic prediction techniques to cognitive radio systems would impact the communication performance of both primary and secondary users. To maintain a trade-off between performance measurements including channel switching rate and call blocking rate of secondary users, interference on primary users and spectrum reuse efficiency, secondary users can set a threshold \(P_{th} (P_{th} \in [0, 1])\) to determine whether to use a channel. If the probability \(P_i (i = 1, 2)\) is not less than this threshold, i.e.,

\[
P_i \geq P_{th}
\]

then, a secondary user can proceed to use the channel. Otherwise, it would abandon using the channel.

The probability threshold \(P_{th}\) in (14) is determined by the requirements on performance measurements, whereas performance measurements are affected by the probabilities of false alarm and missed detection. Here, false alarm refers to the condition that a secondary user judges that primary users would appear whereas, actually, no primary user appears; missed detection refers to the condition that a secondary user judges that no primary user would appear whereas, actually, primary users appear.

To balance the performance in terms of the call blocking rate of secondary users, spectrum reuse efficiency and the interference on primary users, the probability threshold should be selected such that the occurrence rate of false alarm is equal to that of missed detection (Li & Zekavat, 2009).

Defining \(P_{avg}\) as:

\[
P_{avg} = \frac{\sum P_i}{N},
\]

where, \(\sum P_i\) is the summation of the evaluated probabilities of channel availability over all \(N\) decisions (assuming, for one channel, a secondary user makes \(N\) decisions on channel availability).

Based on the discussion in (Li & Zekavat, 2009),

\[
\mathbb{P}\{P_i \geq P_{th}\} = P_{avg}.
\]

Considering \(P_i\) is uniformly distributed between \((1 - \alpha)P_{avg}\) and \((1 + \alpha)P_{avg}\) \((0 \leq \alpha \leq 1, \alpha\) is used to vary the range\), according to (16),

\[
P_{th} = \left(\frac{\sum P_i}{N}\right)(1 + \alpha - 2\alpha)\left(\frac{\sum P_i}{N}\right), 0 \leq \alpha \leq 1.
\]
where, \( \sum P_i \) is introduced in (15).

Taking advantage of the intelligence of cognitive radio, secondary users can dynamically select an appropriate \( \alpha \) based on the evaluation results of probability of channel availability.

If \( P_i \) is assumed to be uniformly distributed between \( P_i^{(\text{max})} \) and \( P_i^{(\text{min})} \), secondary users can record the \( P_i^{(\text{max})} \) and \( P_i^{(\text{min})} \) for each time interval. Then, according to (16), \( P_{th} \) corresponds to:

\[
P_{th} = P_i^{(\text{max})} - \frac{1}{N} \left( P_i^{(\text{max})} - P_i^{(\text{min})} \right).
\]

(18)

Within a time interval, for one channel, the secondary user can set the threshold \( P_{th} \) according to (17) or (18) by using the past evaluated probability results. Note that, within a time interval, the selection of \( P_{th} \) is impacted by the number of primary user call arrivals. Therefore, in different time intervals, \( P_{th} \) should be set dynamically.

The impact of traffic prediction on communication performance is evaluated in (Li & Zekavat, 2009) based on simulations.

5. Probability of secondary users’ successful transmissions

Before starting a transmission, a secondary user needs to check the channel availability. Here, if a channel is available, it means that currently the channel is idle, and the probability that the channel would not be occupied by primary users within \( T \) (\( T \) is the average call holding time of the secondary user) is not less than the probability threshold. Thus, at time \( t \), the probability \( P_a \) that at least one channel is available to secondary users corresponds to (considering total \( m \) channels are licensed to primary users):

\[
P_a = 1 - \prod_{k=1}^{m} \left[ 1 - P_c^{(k)} \cdot \mathbb{P}\{P_i^{(k)} \geq P_{th}^{(k)}\} \right].
\]

(19)

where, \( P_c^{(k)} \) is the probability that channel \( k \) is idle at time \( t \); \( P_i^{(k)} \) is the evaluated probability that channel \( k \) would not be occupied by primary users within \( T \), and \( P_{th}^{(k)} \) is the probability threshold set for channel \( k \). When the hourly number of call arrivals, \( \lambda_n \), increases, \( P_c^{(k)} \) would decrease. If, at time \( t \), a secondary user has a transmission request, then, the probability \( P_{br} \) that this request would be blocked corresponds to:

\[
P_{br} = 1 - P_a = \prod_{k=1}^{m} \left[ 1 - P_c^{(k)} \cdot \mathbb{P}\{P_i^{(k)} \geq P_{th}^{(k)}\} \right].
\]

(20)

To examine the probability that there is a successful secondary user’s transmission at time \( t \), \( P_b \) is assumed to be the probability that there is a secondary user’s transmission request at time \( t \). Then, the probability \( P_s \) that, at time \( t \), there is a successful secondary user’s transmission corresponds to:

\[
P_s = P_a \cdot P_b = \left\{ 1 - \prod_{k=1}^{m} \left[ 1 - P_c^{(k)} \cdot \mathbb{P}\{P_i^{(k)} \geq P_{th}^{(k)}\} \right] \right\} \cdot P_b.
\]

(21)

Generally, the occupancy of the channels is correlated. Here, for simplicity, \( P_c \) is assumed equal for all channels, i.e., \( P_c^{(1)} = P_c^{(2)} = \cdots = P_c^{(m)} \). With another assumption that \( \mathbb{P}\{P_i^{(k)} \geq \)
According to (21), the probability $P_s$ with respect to different $P_b$ and $P_c$ is sketched in Fig. 4 ($m = 8$). The probability $P_s$ is low when either $P_b$ or $P_c$ is low (if the competition among secondary users in using the channel is considered, $P_s$ would be lower). This is due to the fact that, when $P_b$ is low, at time $t$, even if channels are available, nonetheless, secondary users may not have transmission requests; when $P_c$ is low, i.e., $\lambda_n$ is high, the transmission requests from secondary users would mostly be blocked (this can be illustrated by Fig. 5). In other words, requesting channels from secondary users and channels becoming available do not occur simultaneously. This implies that, in some situations, even if channels which are licensed to primary users are underutilized statistically, it is still hard for secondary users to obtain available channels because channel access by primary users and channel request from secondary users are both random processes.

6. Cooperative Prediction

To improve the accuracy of call arrival prediction, a secondary user may collect the prediction results of primary users’ call arrivals from other secondary users, and obtain the average value of the predicted call arrivals. This is called cooperative prediction. Secondary users which are involved in the cooperative prediction should be able to monitor the same channel(s). Furthermore, a common channel (mostly, with low bandwidth) is required for those secondary users to share the prediction results. After collecting $M$ prediction results of $\lambda_{n+1}$ from other $M$ secondary users, the secondary user may calculate the average value $\bar{\lambda}_{n+1}$ to evaluate the
Fig. 4. Probability that there is a successful secondary user’s transmission at time $t$.

Fig. 5. Channel utilization and secondary users’ channel requests.

The probability of channel availability, $\bar{\lambda}_{n+1}$, can be calculated as:

$$\bar{\lambda}_{n+1} = \frac{\sum_{i=1}^{M+1} \hat{\lambda}_{n+1}^{(i)}}{M+1},$$

where, $\hat{\lambda}_{n+1}^{(i)}$ is the prediction result of $\lambda_{n+1}$ from secondary user $i$.

If the average call holding time $\bar{T}$ is the same for those involved secondary users, another cooperation technique can be employed (Unnikrishnan & Veeravalli, 2008); (Aalo & Viswanathan, 1992). Considering two hypotheses $H_1$ and $H_0$:

$$\left\{ \begin{array}{l} H_1 : P_i \geq P_{th} \\ H_0 : P_i < P_{th} \end{array} \right.$$

for secondary user $k$, the decision $U_k$ corresponds to:

$$U_k = \left\{ \begin{array}{ll} 1, & \text{if } H_1 \text{ is true} \\ -1, & \text{if } H_0 \text{ is true} \end{array} \right.$$
After receiving decisions in the form of $U_i$ from other $K$ secondary users, one secondary user can calculate the summation of all $U_i$’s. If the summation result is greater than 0, i.e.,

$$\sum_{i=1}^{K+1} U_i > 0,$$  

then, the secondary user can proceed to use the channel. Otherwise, it should abandon using the channel. Fig. 6 illustrates the whole process, in which $w_1, \cdots, w_K$ are the weights that can be simply considered as one or tuned based on field data.

7. Traffic Prediction based Spectrum Sharing

Cognitive Radio (CR) techniques enable the inter-vendor and user-central spectrum sharing (Zekavat & Li, 2005); (Li & Zekavat, 2007). The approach of finding available spectrum based on traffic pattern prediction can be applied to different cognitive radio based spectrum sharing techniques.

7.1 Spectrum Sharing across Multiple Service Providers

In business zones, for an infrastructure (e.g., base station) of a service provider, when the number of active users is greater than the maximal number of users that it can accommodate (i.e., the infrastructure is in overloaded status due to channel scarcity), the incoming users would be blocked. However, at the same time, nearby infrastructures of other service providers might be in the underloaded status. Available channels licensed to these service providers or other organizations can be used by the overloaded infrastructure (assuming multiple service providers coexist in the same area). The difference in traffic load across multiple service providers might be due to the diversity of services that service providers offer to their customers, e.g., mobile communication, wireless Internet for laptops, etc.

Generally, there are two approaches for the overloaded infrastructure to find available channels which are licensed to other service providers: (a) the overloaded infrastructure asks neighboring service providers for available channels; (b) the overloaded infrastructure is equipped with CR, and it senses the available channels within its coverage area. However, if the cell radii of service providers involved are different, and/or the infrastructures of different service providers are not located at the same positions, then, both approaches could lead to co-channel interference.

To reduce the co-channel interference and remove the need of equipping CR in each infrastructure of different service providers, a CR network consisting of multiple fixed CR nodes can be deployed to support the spectrum sharing. These CR nodes are distributed regularly within the area of interest, and each CR node is a dedicated node that senses the surrounding environment and monitors the channel usage within its sensing range. Those channels might
be licensed to different service providers. To avoid co-channel interference, CR nodes which monitor the channel usage in the cell of interest and the corresponding co-channel cells cooperate to provide channel availability information for the overloaded infrastructures of service providers. CR nodes are connected together via wire or they communicate wirelessly to form a network. The CR network can be called a spectrum management network, and coexists with the wireless networks operated by different service providers (see Fig. 7). Thus, when one or more infrastructures of a service provider are overloaded, they request the channel availability information from those surrounding CR nodes. Then, overloaded infrastructures process the information to select the optimum channels based on the channel associated metrics such as interference level and the probability of channel being available for a certain time duration. A CR node would determine channel $i$ is available if: 1) The instantaneous RF (Radio Frequency) energy (plus noise) in this channel, $I^{(i)}_{int}$, is less than the tolerable interference limit, $I^{(i)}_{tol}$, of this channel, i.e., $I^{(i)}_{int} < I^{(i)}_{tol}$, and, 2) The probability $P^{(i)}$ that this channel would be available (not occupied by its licensed users) for a given time period is not less than a threshold $P^{(i)}_{th}$, i.e., $P^{(i)} \geq P^{(i)}_{th}$. $P^{(i)}$ is evaluated by CR nodes using the discussed traffic pattern prediction technique. Thus, it reduces the probability that the selected channels are claimed by their licensed users when these channels are being used by the overloaded infrastructure.

8. Conclusion

In this chapter, traffic models and traffic prediction techniques which can be implemented in secondary (cognitive) users are discussed. Methods are introduced for secondary users to evaluate the probability of channel being available for a given time period in different situations. Comparing the evaluated probability of channel availability with a threshold, secondary users can determine whether to use a channel. The threshold maintains a trade-off between the channel switching rate and call blocking rate of secondary users, the interference on primary users and spectrum reuse efficiency. In addition, the probability that there is a successful secondary user’s transmission at one time instance is examined, and cooperative prediction techniques are briefly introduced. Finally, the application of the traffic pattern prediction technique in spectrum sharing is discussed.
9. References


Cognitive radio is a hot research area for future wireless communications in the recent years. In order to increase the spectrum utilization, cognitive radio makes it possible for unlicensed users to access the spectrum unoccupied by licensed users. Cognitive radio lets the equipments more intelligent to communicate with each other in a spectrum-aware manner and provide a new approach for the co-existence of multiple wireless systems. The goal of this book is to provide highlights of the current research topics in the field of cognitive radio systems. The book consists of 17 chapters, addressing various problems in cognitive radio systems.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following: