Overload Prediction Based on Delay in Wireless OFDMA Systems

E. O. Lucena, F. R. M. Lima, W. C. Freitas Jr and F. R. P. Cavalcanti
Federal University of Ceará - UFC, Wireless Telecommunications Research Group - GTEL
CP 6005, Campus do Pici, 60455-760, Fortaleza-CE, Brazil
{evilasio,rafaelm,walter,rodrigo}@gtel.ufc.br

***Abstract*—In this work we deal with Congestion Control (CC) strategies to protect the Quality of Service (QoS) of Real Time (RT) services in Orthogonal Frequency Division Multiple Access (OFDMA)-based and packet-switched networks such as Long Term Evolution (LTE) and Worldwide Interoperability for Microwave Access (WiMAX). Specifically, the first contribution of this paper is the adaptation of a CC strategy previously proposed for High-Speed Downlink Packet Access (HSDPA) to these modern systems. This CC strategy comprises in a coordinated manner the functionalities scheduling, Admission Control (AC) and Load Control (LC). The second contribution is the proposal of a new feature to be added to this CC strategy in order to allow for an early detection of overload situations based on the packet delay of RT flows. In the results we show that our proposed overload prediction based on delay can efficiently prevent QoS degradation of RT flows.**

***Index Terms*—Congestion Control, Overload Prediction, RT Services, Delay, OFDMA.**

I. Introduction

The mobile networks are continuously evolving in order to support more users, to achieve higher data rates, and to provide new (multimedia) services. When delivery requirements are concerned, the services are categorized as Non-Real Time (NRT) or Real Time (RT). RT services require a short time response between the communicating parts and, in general, impose strict requirements regarding packet delay and jitter. Voice over IP (VoIP) and online games are examples of services of this class.

Current IP networks are designed for best-effort services lacking stringent Quality of Service (QoS) control. Congestion is inevitable in these networks and may result in packet loss, delay, and delay jitter, which directly impact on the QoS of RT applications. Thus, the current IP network architecture must be enhanced by some QoS guaranteeing mechanisms in order to ensure QoS of RT applications [1].

One strategy to guarantee the desired QoS for RT services is the utilization of prioritization among services such as scheduling [2]. However, when the system is near to be overloaded scheduling alone cannot guarantee QoS for RT services. Overload situations can cause variations in cell load and in the QoS experienced by the users. In these situations, the prioritization should be applied in a broader sense by means of Congestion Control (CC) algorithms. The work [3] proposed a QoS-driven adaptive CC framework that joins the functionalities of scheduling, Admission Control (AC) and Load Control (LC). The objective of that framework is to guarantee the QoS of RT flows in the High-Speed Downlink Packet Access (HSDPA) system in multiservice scenarios.

In this work we propose two main contributions. The first one is the generalization of the CC framework proposed in [3] to work with multiple subcarriers to be considered for networks employing Orthogonal Frequency Division Multiple Access (OFDMA) in the downlink. The generalized framework is called hereafter *AdaptiveCC*. The second one is a new feature based on delay to be added to the generalized framework to predicts an overload situation. In order to prevent high peaks of Frame Erasure Rates (FERs), the generalized framework with the feature of overload prediction based on delay is called hereafter *Delay-based Prediction*.

The remainder of this document is organized as follows: in section II we show the *Adaptive CC* framework. In section III we present the motivation and formulation of the *Delay-based Prediction* framework and in section IV the performance evaluation and its results in a case study with VoIP and World Wide Web (WWW) services. Finally the main conclusions and perspectives are summarized in section V.

II. Adaptive Congestion Control Framework

The *Adaptive CC* framework for OFDMA systems comprises in a coordinated manner the operation of AC, scheduling and LC algorithms. In this section we review the AC and LC algorithms and show the main modifications in the scheduling algorithm.

A. Admission control

In this work, a Session Admission Control (SAC) scheme is employed to guarantee the quality of a RT service in a mix with other services. The SAC algorithm considers delay as the resource to be shared among flows in the system. In order to do that, the packet delays of the RT traffic are regularly measured and filtered by an Attack-Decay (AD) Filter. Each transmitted or discarded packet provides a delay sample that feeds the AD filter.

There are two admission thresholds depending on the service type: D_{RT}^{th} for the RT service and D_{Other}^{th} for other low-priority services. Therefore, when a new RT flow tries to

This work was supported by the Research and Development Center, Ericsson Telecomunicações S.A.,Brazil, under EDB/UFC.22 Technical Cooperation Contract.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

access the system, the SAC algorithm will check if the filtered delay calculated by the AD filter, represented hereafter by D_{RT}, is greater or lower than D_{RT}^{th}. If $D_{RT} > D_{RT}^{th}$ the new flow is rejected; otherwise the flow is admitted. The procedure is the same if the new flow is of an NRT service.

According to the admission thresholds, a service can be more prioritized than the others. Although in [4] these admission thresholds are fixed, the main idea in the *Adaptive CC* framework is to adapt them according to the congestion status of an RT service.

In order to prioritize a service, we should introduce a new variable: the SAC priority margin, α. We define this priority margin (in decibel) as

$$\alpha[k] = 10\log_{10}\left(\frac{D_{Other}^{th}[k]}{D_{RT}^{th}}\right), \quad (1)$$

where $\alpha[k]$ is the SAC priority margin at Transmission Time Interval (TTI) k.

As we can see, this variable defines a ratio between the admission threshold of two services and it can define a difference in their priorities once D_{RT}^{th} can assume a fixed value and α can assume different values at different TTIs. The LC algorithm is responsible for the adaptation of α and so to establish different priorities between the services as it will be discussed in section II-C. More details about SAC can be found in [3], [4].

B. Scheduling

The main change made in this work in order to generalize the framework developed in [3] to a network that works with multiple subcarriers is in the scheduling algorithm. Before this work, [3] used the Weighted Proportional Fair (WPF) [2]. With WPF, the (single) scheduled flow is the one with highest priority. The priority of a flow j at TTI k is given by

$$p_j[k] = w_j[k].\left(\frac{r_j[k]}{t_j[k]}\right), \quad (2)$$

where to flow j at TTI k $w_j[k]$ represents a service-dependent weight, $r_j[k]$ is the supported data rate and $t_j[k]$ is the filtered data rate according to the channel state. The filtered data rate t provides a history of the allocated data rates in the past. On the one hand, if the flow is from an RT session, $w_j[k]$ is set to W_{RT},and on the other hand, if the flow is from another service the weight is equal to W_{Other}. Therefore, by setting different values to these weights some prioritization among the services can be accomplished. For this reason, despite in [2] W_{RT} and W_{Other} are fixed, the *Adaptive CC* framework adapts W_{Other} to control the congestion in the RT service.

In an OFDMA-based system, the frequency diversity can be exploited by adding another dimension in that prioritization. Therefore, we have adopted the Weighted Multicarrier Proportional Fair (WMPF) scheduler [5] that is a natural generalization of WPF. The prioritization in WMPF is given by

$$p_{j,n}[k] = w_j[k].\left(\frac{r_{j,n}[k]}{t_j[k]}\right), \quad (3)$$

where $p_{j,n}[k]$ is the priority of j in subcarrier n at TTI k and $r_{j,n}[k]$ is the supported data rate of j in subcarrier n at TTI k (according to the channel state of subcarrier n). From the priorities $p_{j,n}[k]$ we can build a priority matrix. The flow selection consists in assigning the pair flow-subcarrier corresponding to the largest entry in the priority matrix. In this way, multiple flows can be scheduled simultaneoulsy with potentially higher data rates.

After defining the priority $p_{j,n}[k]$, we should introduce a new variable: the WMPF priority margin, $\beta[k]$. We define this priority margin (in dB) as

$$\beta[k] = 10\log_{10}\left(\frac{W_{Other}[k]}{W_{RT}}\right), \quad (4)$$

where W_{RT} and $W_{Other}[k]$ represent service-dependent weights of the RT flow and other service, respectively, and $\beta[k]$ is the value of β at TTI k.

As we can see, this variable defines a ratio between the weights of two services and it can define a difference in their priorities once W_{RT} can assume a fixed value and β can assume different values at different TTIs. The LC algorithm is responsible for the adaptation of β and so to establish different priorities between the services as it will be discussed in section II-C.

C. Load control

Considering that D_{RT}^{th} and W_{RT} are fixed reference values, the dynamic adaptation of the priority margins α and β can control the prioritization of the RT service over other services. The main idea of the LC algorithm is to adapt these priority margins according to the QoS of the ongoing sessions of the high priority service. If the QoS of the sessions of the RT service is not being fulfilled, the LC algorithm decreases the priority margins. As a consequence, the sessions of the RT service will be scheduled more often and the system will decrease the number of admitted sessions of other services in order to protect the ongoing sessions of the RT service. The adaptation of the priority margin at each TTI k is calculated by means of a RT service metric. Without loss of generality, we consider the FER as the main performance metric of RT services. However, another metric could be used instead. The adaptation of the priority margin $\alpha[k]$ is given as follows:

$$\alpha\left[k\right] = \min\left\{\max\left\{\alpha_{\min}, \alpha\left[k-1\right] - \sigma_\alpha \cdot e\left[k\right]\right\}, \alpha_{\max}\right\}, \quad (5)$$

where $e[k]$ is given by

$$e\left[k\right] = FER_{RT}^{filt}\left[k\right] - FER_{RT}^{target}. \quad (6)$$

The FER considers a ratio of number of lost frames (or packets) and the total number of generated packets. FER_{RT}^{filt} is the filtered FER in the last control interval and FER_{RT}^{target}

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

is the target FER to experience a good QoS. The filtered FER $FER_{RT}^{filt}[k]$ is obtained by applying a Simple Exponential Smoothing (SES) filter to the time series comprised by the average FER in each TTI [6]. It is important to observe that $\beta[k]$ is adapted in the same way as $\alpha[k]$ where $\alpha_{\min}$, $\alpha_{\max}$ and σ_α are replaced by $\beta_{\min}$, $\beta_{\max}$ and σ_β, respectively. $\alpha_{\min}$, $\alpha_{\max}$, $\beta_{\min}$ and $\beta_{\max}$ are the minimum and maximum values in dB of the $\alpha[k]$ and $\beta[k]$ parameters, respectively. The fixed parameters σ_α and σ_β control the adaptation speed of the priority margins α and β, respectively. For further details about the LC strategy see [3], [7].

III. Overload Prediction Based on Delay

The second contribution proposed in this work is a new feature to be added to this CC strategy in order to allow for an early detection of overload situations based on the packet delay of RT flows in a framework called *Delay-based Prediction*. In this section we present the motivation in measuring FER to obtain the overload prediction based on delay. In addition, we present the formulation and performance evaluation of *Delay-based Prediction*.

A. Motivation

The *Adaptive CC* presented in the previous section is capable of protecting the QoS of the RT services by monitoring a service specific metric. The prioritization among services in scheduling, AC and LC are changed in order to *react* when an overload situation is detected. However, as the reaction of the *Adaptive CC* takes place when the overload already exist, the system will get back to normal load conditions only after a certain period. Within this period, RT sessions can be compromised due to poor QoS experience. Therefore, we believe that the QoS of RT sessions could be even protected if a *predictive* capacity would be added to the *Adaptive CC*.

The packet delay plays an important role in the perceived QoS of RT services. Usually, the packets of RT services have stringent delay requirements. In case the packet delay deadlines are violated, these packets are usually discarded by upper protocol layers and the overall QoS experienced by the end user is degraded. Therefore, if in average the packet delays of the active RT flows are increasing, this is a strong indication that packet discard will happen. With this in mind, we propose the addition of a new feature to the *adaptive CC* that is the capability of predicting and avoiding overload situations by using measurements of the packet delays of RT flows.

B. Formulation

The increasing delay information can be added to the priority margins α and β so that scheduling weigths and admission thresholds can start their updates as the delay increases and gets close a threshold before the packet loss occurs and the FER increases.

The variable Y works adding a value to the FER_{RT}^{filt}. Hence, the Equation (6) can be modified and yields

$$e[k] = \left(FER_{RT}^{filt}[k] + Y\right) - FER_{RT}^{target}. \quad (7)$$

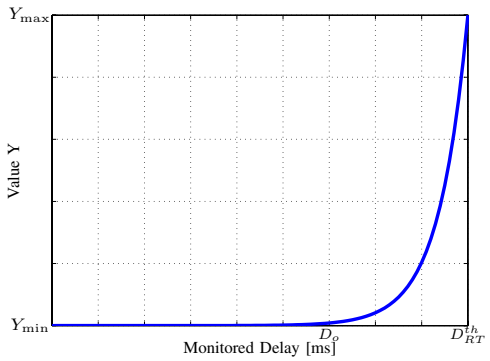

Figure 1: Behavior of the Delay Prediction Variable Y.

Figure 1 illustrates the behavior of the delay prediction variable Y that is calculated as:

$$Y = Y_{\min} - M + M \cdot \exp\left(\frac{D_{VoIP}^{filt}}{D_{RT}^{th}} \cdot \ln\left(\frac{Y_{\max} - Y_{\min} + M}{M}\right)\right), \quad (8)$$

where M is a constant responsible for the slope of the exponential curve, D_{VoIP}^{filt} is the filtered delay experienced by all packets from all users in the cell, D_{RT}^{th} is the RT SAC delay threshold, Y is an adjustment factor to the filtered FER, $Y_{\min}$ and $Y_{\max}$ are fixed parameters that indicate the minimum and the maximum values of Y, respectively.

It is worth noticing that Y increases only when the monitored delay is higher than a threshold D_0, and so $e[k]$ also increases, simulating a higher FER. Thus, we force a reaction of the system before the increase of the FER. Besides, the larger $e[k]$ is, the faster the *Delay-based Prediction* framework works.

IV. Performance Evaluation

In this section, we present a performance evaluation of the *Delay-based Prediction* framework compared to a framework without overload prediction called *Adaptive CC* and to a reference framework called *Non Adaptive CC*, where no CC solution is applied, i.e., in which the WMPF and SAC priority margins are not adapted. Details about the simulation tool and the main parameters used to obtain the results are presented in section IV-1. In section IV-2 we define the performance metrics used in this study. Finally, in section IV-3 we show and analyze the simulation results.

1) Simulation parameters: The simulation tool models the main aspects of a single-cell OFDMA system[1]. The results can be obtained with different congestion levels that are related to the interval time between the arrival of different flows in the system. Concerning services we consider a mixed traffic scenario with VoIP as the RT service and WWW as the NRT one. The flows arrive in the system according to a Poisson

[1]Albeit the adaptive CC framework present in this work can also be performed in multiple-cell scenarios.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

distribution. Following the new flow arrival some propagation parameters are calculated based on its position inside the cell such as its path loss, its shadowing and its path gain. The propagation conditions change once the position of the flow also changes. We assume that the data symbols are independently modulated and transmitted over a high number of closely spaced orthogonal subcarriers. The modulation schemes Quadri-Phase Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (QAM), and 64 QAM are available and are chosen depending on the flow Signal-to-Noise Ratio (SNR) that varies with the propagation conditions. A session is finished when a flow has no more packets to transmit. The simulation ends when a certain number of sessions is achieved.

Table I: Main parameters of the simulation tool.

Parameter	Value	Unit
Cell radius	500	m
Minimum distance from Base Station (BS)	100	m
Maximum BS power	5	W
Number of Subcarriers	200	-
Subcarrier spacing	15	kHz
Carrier frequency	2	GHz
Fast Fading speed	3	km/h
Path loss [8]	$128 + 37.6 \cdot \log_{10}(d)$	dB
Standard deviation of lognormal shadow fading	8	dB
Small-scale fading	multiple-path Rayleigh	-
VoIP SAC delay threshold (D^{th}_{VoIP})	100	ms
VoIP WMPF priority weight (W^{prio}_{VoIP})	1	-
Time basis for adaptation of α	100	ms
Time basis for adaptation of β	1	ms
Maximum value of α and β ($\alpha_{\max}$, $\beta_{\max}$)	0	dB
Minimum value of α and β ($\alpha_{\min}$, $\beta_{\min}$)	-10	dB
SAC step size (σ_α)	0.5	dB
WMPF step size (σ_β)	0.5	dB
VoIP satisfaction requirement	95	%
WWW satisfaction requirement	90	%
VoIP FER threshold	1	%
WWW throughput threshold	128	kbps
Maximum value of adjustment factor to FER^{filt}_{VoIP} ($Y_{\max}$)	0.05	–
Minimum value of adjustment factor to FER^{filt}_{VoIP} ($Y_{\min}$)	0	–
Constant responsible for the slope of the exponential curve M	$3 \cdot 10^{-8}$	–
Time basis for adaptation of Y	100	ms

2) Performance metrics: For all present results, we define the offered load as the mean flow arrival rate (in number of flows per second) in the system. This is an input parameter to the Poisson processes used to model flow arrivals.

An important metric used in this study is the satisfaction ratio of a service. For a VoIP flow, the satisfaction is reached when it was not blocked by the AC functionality and its experienced FER is equal to or lower than 1% [9]. For the WWW service, the flow is satisfied when it was not blocked and its average throughput assumes at least the rate requirement value of 128 kbps.

3) Results: We start this section illustrating in Figure 2 the time variation of the filtered FER (FER^{filt}_{VoIP}) and the filtered delay (D^{filt}_{VoIP}). First of all, we can see in this figure that there is a strong correlation between the packet delays and the QoS of VoIP flows represented by the FER. In fact, peaks of the filtered FER are usually preceded by high values of filtered packet delays for the three presented frameworks.

When the performance of the frameworks are concerned, Figure 2 also provides important insights. We can see that the three presented frameworks succeed in controlling an overload situation characterized by $FER^{filt}_{VoIP} > 1\%$ in this case. However, our proposed *Delay-based Prediction* framework presents an important difference compared to the two other frameworks: the overload predictive capacity. The filtered FER for the *Non Adaptive CC* and *Adaptive CC* frameworks is controlled only when the filtered FER is higher than the target FER. As we can see in this figure, although the filtered FER presents an increase when the filtered delay is high with the *Delay-based Prediction* framework, the filtered FER does not overcome the target FER. This is achieved by the use of packet delay measurements in order to preview overload in the VoIP service.

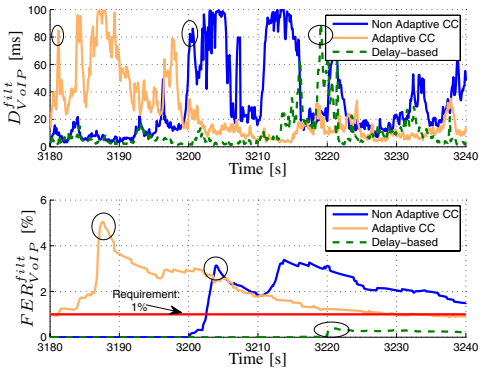

Figure 2: FER^{filt}_{VoIP} and D^{filt}_{VoIP} for VoIP users in three different frameworks: *Non adaptive CC*, *Adaptive CC* and *Delay-based Prediction* at load of 1.375 Users/s for mix 25% of VoIP and 75% of WWW flows.

In order to complement the information provided by the Figure 2 we show in the Table II the performance gains of the *Delay-based Prediction* compared to the two other frameworks in reducing the mean FER^{filt}_{VoIP} for different mix of services. This result shows that our proposed *Delay-based Prediction* is able to control the FER in overload conditions improving QoS experienced for the end user.

Table II: Performance gains of *Delay-based Prediction* in reducing the mean FER^{filt}_{VoIP}.

Mix of Service	*Adaptive CC*	*Non Adaptive CC*
Mix 25% VoIP 75% WWW	21.64%	79.23%
Mix 50% VoIP 50% WWW	8.87%	76.49%
Mix 75% VoIP 25% WWW	51.74%	70.74%

In Figures 3, 4 and 5 we present the satisfaction ratio for different mixes of services. We can notice that higher is the percentage of VoIP flows the greater is the performance gain for VoIP if we compare the *Delay-based Prediction* framework and the *Adaptive CC*. The reason for this is that the WWW demands a lot of resources, so the lower the number of WWW flows the greater will be the performance gain in VoIP service. The performance gain obtained in *Delay-based Prediction*

978-1-4244-5638-3/10/$26.00 ©2010 IEEE

framework occurs because the system monitors the delay and so it is possible to predict that a VoIP packet can be lost before it happens; hence the system can react before FER builds up.

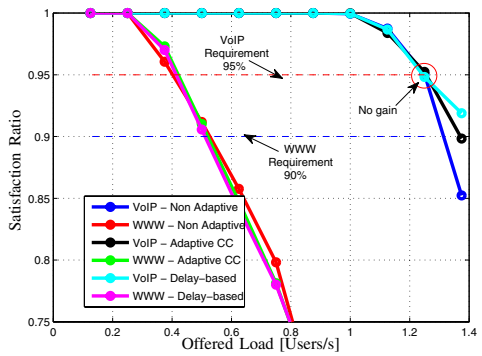

Figure 3: Satisfaction for three different frameworks: *Non adaptive CC*, *Adaptive CC* and *Delay-based Prediction* for mix 25% of VoIP and 75% of WWW flows.

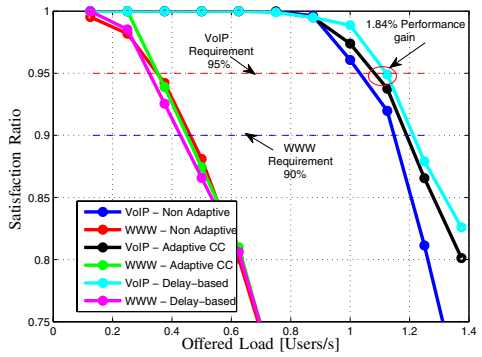

Figure 4: Satisfaction for three different frameworks: *Non adaptive CC*, *Adaptive CC* and *Delay-based Prediction* for mix 50% of VoIP and 50% of WWW flows.

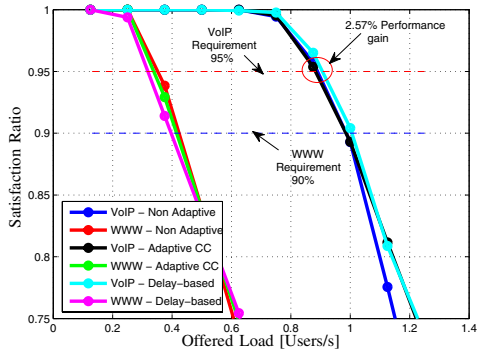

Figure 5: Satisfaction for three different frameworks: *Non adaptive CC*, *Adaptive CC* and *Delay-based Prediction* for mix 75% of VoIP and 25% of WWW flows.

The satisfaction results have shown that VoIP service has in general a satisfaction level greater than that obtained for the WWW one. Due to VoIP packets characteristics and to the higher resource granularity found in OFDMA, which provides frequency chunks for resource allocation, VoIP flows need fewer resources to empty their buffer. For this reason the WWW service is considered the restricting one that limits the joint capacity. Nevertheless, observing Figures 3, 4 and 5 we can conclude that *Delay-based Prediction* imposes only a small performance degradation to the WWW service in order to guarantee the QoS fulfillment of VoIP.

V. Conclusion

In this work we presented two contributions to protect the QoS of RT services in a mixed traffic scenario. The first contribution is the generalization of the CC framework proposed in [3] for networks employing OFDMA in the downlink. In the second contribution we proposed a new feature to be added to the generalized framework. The *Delay-based Prediction* framework besides guaranteeing the QoS fulfillment of a RT service also prevent high peaks of FERs by using the packet-delay prediction capability. By analyzing simulation results, we can conclude that the overload prediction based on delay can efficiently avoid the increase of FER before it really builds up and prevent QoS degradation of RT flows imposing only a small performance degradation to the WWW service.

References

[1] X. Chen, C. Wang, D. Xuan, Z. Li, Y. Min, and W. Zhao, "Survey on QoS Management of VoIP," October 2003, pp. 69 – 77.

[2] A. R. Braga, E. B. Rodrigues, and F. R. P. Cavalcanti, "Packet Scheduling for VoIP over HSDPA in Mixed Traffic Scenarios," in *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on*, Helsinki, September 2006, pp. 1–5.

[3] E. B. Rodrigues and F. R. P. Cavalcanti, "QoS-Driven Adaptive Congestion Control for Voice over IP in Multiservice Wireless Cellular Networks," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 100–107, January 2008.

[4] A. R. Braga, S. Wanstedt, and M. Ericson, "Admission Control for VoIP over HSDPA in a Mixed Traffic Scenario," in *Telecommunications Symposium, 2006 International*, Fortaleza, Ceara, September 2006, pp. 71–76.

[5] Y. C. L. Yanhui, W. Chunming and T. Guangxin, "Downlink Scheduling and Radio Resource Allocation in Adaptive OFDMA Wireless Communication Systems for User-Individual QoS," in *Transactions on Engineering, Computing and Technology - World Enformatika Society*, no. 2, March 2006, pp. 426–440.

[6] R. E. Goot, U.Mahalab, and R. Cohen, "Nonlinear Exponential Smoothing (NLES) Algorithm for Noise Filtering and Edge Preservation," in *HAIT Journal of Science and Engineering*, vol. 2, May 2005.

[7] E. B. Rodrigues, F. R. M. Lima, and F. R. P. Cavalcanti, "Load Control for VoIP over HSDPA in Mixed Traffic Scenarios," in *Personal Indoor and Mobile Radio Communications*, Athens, Greece, September 2007.

[8] 3GPP, "Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS," UMTS/ETSI, Tech. Rep. TR 101.112 v3.2.0, April 1998.

[9] ——, "Performance Characterization of the Adaptive Multi-Rate Speech Codec," 3^{rd} Generation Partnership Project, Tech. Rep. TS 25.975 V6.0.0 - Release 6, December 2004.

978-1-4244-5638-3/10/$26.00 ©2010 IEEE