UDC 654.165
doi:10.15217/issn1684-8853.2015.2.90

AN EFFICIENT CROSS-LAYER AWARE MAPPING OF VOIP CALLS IN WIRELESS OFDMA SYSTEMS Part I. Problem description and channel tracking

Y. Ben-Shimol[a], *PhD, Electrical Engineering, Professor, benshimo@bgu.ac.il*
I. Kitroser[a], *PhD, Electrical Engineering, kitroser@bgu.ac.il*
[a]*Ben-Gurion University of the Negev, POB 653, 1, Ben Gurion St., Beer Sheva, 74105, Israel*

Purpose: *This work addresses the problem of efficient broadcasting of resource allocations descriptors for VoIP traffic in mobile OFDMA-based wireless systems.* ***Methods:*** *We present the problem of mapping overhead and show that it can be substantially reduced by using semi-persistent allocations and by taking advantage of the periodicity of VoIP frames, generated by a multi-phase vocoder. To handle the impact of mobility on the characteristics of the wireless channel we utilize a cross-layer decision approach that tracks the channel quality and predicts the expected mobile user behavior such that the system may under-react to channel changes in some cases.* ***Results:*** *We explicitly show how the variability of the wireless channel can be tracked by using a Markovian model with a set of discrete states. By estimating both state- and transition probabilities of the multi-state Markovian model, we underline the foundations for a cross-layer decision algorithm that is able to overlook short transitions in channel states, without changing the modulation and/or the coding schemes. The main advantage of this approach is the ability to support multiple codecs, or a single codec with different operational modes, both result in different packet sizes and different periods.* ***Practical relevance:*** *The proposed channel tracking mechanism is simple enough to be implemented in the base station of any practical OFDMA-based system using the WiMAX or LTE technology.*

Keywords *— Cross-Layer Design and Optimization, Resource Allocation and Interference Management, Mobile Multimedia Technology, 3.5G and 4G Technologies.*

Introduction

An important application in future IP based wireless communication networks is voice over IP (VoIP), which allows real-time voice calls between mobile stations (MSs). Most VoIP applications generate traffic of fixed size packets at a constant rate. Therefore, the traffic of a single VoIP session is classified as constant bit rate (CBR) traffic type and requires a fixed number of uplink (UL) resources on a periodic basis. This suggests that with a naive resource allocation scheme the uplink map (UL-MAP) includes allocation descriptors (or information elements — IEs) for VoIP streams, thus generating overheads at a constant rate. A typical voice call lasts several minutes on the average, while its packet period is usually up to several tens of milliseconds. This suggests that the total amount of overhead for each VoIP call is usually very large.

The problem of system overhead for CBR-like applications has been examined for the emerging beyond 3G technologies such as WiMAX (IEEE 802.16) [1], IEEE 802.16m-2011 [2], 3GPP long-term evolution (LTE) [3] and 3GPP2 ultra mobile broadband (UMB) [4]. For these systems, the protection of the resource allocation control messages against transmission errors costs, in some scenarios, is more than 50 % of the downlink (DL) radio resources [5]. This extreme overhead could significantly reduce all the potential benefits of OFDMA technologies, and render OFDMA-based systems inefficient and unprofitable.

The concept of persistent allocation, which was introduced in LTE and UMB, and was also adopted by WiMAX [6, 7], uses periodic assignments of VoIP packets hence, reducing the assignment signaling overhead. With persistent allocations, no message passing between the base station (BS) and the MS is required during the allocation periods. This means that the BS knows the requirements (i.e., burst size and period according to the voice flow profile established) of each MS in advance, and each MS knows that as long as no relevant UL-MAP IE has been broadcast, it is allocated the same set of resources. This allows the BS to use persistent allocations and to notify each MS of its allocation only once at the beginning of the session and then only whenever the allocation changes, eliminating almost all other mapping overheads concerned with VoIP calls. The work in [6] with some modifications has been adopted to the IEEE 802.16REV2 of [8, 9] with the final version published as 802.16-2009 [1]. Das, et al. [10] present an algorithm for dynamic and semi-persistent VoIP scheduling in LTE, in which they predict the number of physical resource blocks required for a given number of VoIP users. However, a specific modulation and coding scheme (MCS) selection scheme according to a physical channel condition is not given.

Kitroser and Ben-Shimol [11] presented an efficient mapping algorithm for multiple VoIP coders for WiMAX systems. In general, an efficient mapping algorithm has to consider the parameters of the coders in use and the characteristics of the voice

sessions, such as the average call duration and initiation rate. In each frame the mapping algorithm should change the minimal number of allocations and if a change is inevitable, impose a minimum number of changes in successive frames. A significant system overhead reduction was shown for stationary wireless systems by using the minimum overhead algorithm (MOA) for VoIP mapping. The main idea of MOA is to decrease local and future allocations overhead in each frame by allocating slots in such manner that reduces the potential collision of these slots with allocations in subsequent frames due to different codec periods. Summary on VoIP and its scheduling in OFDMA systems is given in [12, 13] and references therein.

The present paper describes the first part of required extensions to the work presented in [11], in order to support full mobility conditions. In [11] the MOA was designed to efficiently handle multiple codecs with different periods. The algorithm was found to work well for fixed channels but its performance degrades when the channel quality changes through time. The main reason for the performance degradation was the fact that each time the channel changes, the system, which usually employs dynamic MCS selection, must respond with allocation changes due to the rate change. This basically diminishes the main advantage of the semi-fixed allocation concept, since much overhead is again added for control purposes. Any allocation scheme designed only for fixed (stationary) channels suffers from similar degradation in efficiency. To change allocation decisions with relation to channel variations, we show an efficient, yet simple, way to handle mobile channels and MCS variations by using a finite state Markov chain (FSMC) model for approximating and tracking the channel states. This model will serve later (in part II of the present paper) a set of allocation decision algorithms that will complete the cross-layer solution for the problem at hand.

The rest of the paper is formed as follows: Section "Related-Work" discusses the related work of persistent allocation for IEEE 802.16. Section "System Model" presents the system model and notations for the allocation problem in a dynamic channel followed by the formal objective. Section "Channel-Characterization" presents our channel characterization model using FSMC model. A discussion on the logic behind the development of efficient mapping algorithm summarized this part, part II will present a set of efficient mapping algorithms capable to handle mobile users and multi-state vocoders.

Persistent Allocation in IEEE 802.16 Systems

Persistent allocation for IEEE 802.16 systems was proposed by Lu, et al. [6] and introduced to the standard in IEEE 802.16-2009 [1]. It provides specific signaling and error handling for VoIP traffic in order to reduce the inherent overhead caused by such traffic. Besides the work of [6] and concurrently with the persistent allocation and semi-persistent allocation mechanism in other technologies (LTE and UMB, see e.g., [14–16]), several other solutions have been proposed to improve VoIP efficiency for IEEE 802.16 [6, 7, 17–23].

In [17, 18] a persistent allocation scheme with bitmap signaling is described. In [17] the users are grouped into scheduling groups according to channel quality information (using a channel quality index — CQI) obtained from the users. Then, each user receives a unique position within its group. Once a group of users is established, the BS assigns the group a set of shared time-frequency resources and an ordering pattern indicating the order in which the resources are allocated. Finally, bitmap signaling is used to allocate resources in each VoIP frame. This scheme is basically a signaling scheme, however, it fails to specify how to handle channel dynamics when users change CQI and hence need to change scheduling groups. In [18], the authors present a group of scheduling mechanism to overcome the problem of resource holes in the data allocation region of the frame. The MSs are clustered into multiple groups and the resource allocation for individual MSs has some persistency within the group's resources. The grouping and bitmap signaling is then used to accommodate allocation changes. The addressed solution mainly solves the DL allocation problem, since the mechanism of dynamic bitmap signaling is not applicable for UL where the signaling is initiated by the user side. The main advantage of the bitmap approach is that hole filling and dynamic response to codec state is efficient. A disadvantage can be pointed out for the case of static scenarios where channel dynamics are low: in the persistent approach no messages would have been transmitted at all, while in the bitmap approach, the signaling is used in each frame. In addition, it is not clear how this approach can handle multiple codecs when many periods may intersect.

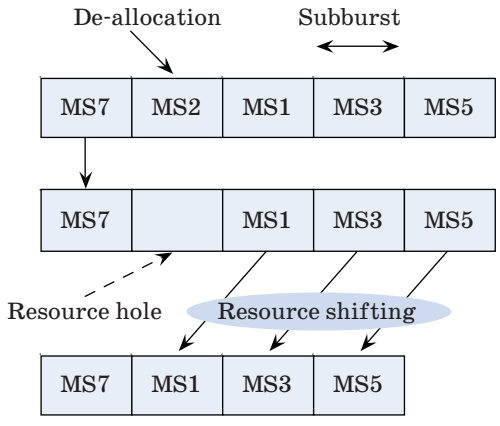
A periodic allocation-mapping scheme is presented in [19] to decrease the size of the allocation map by representing the allocated resources by only the MCS level. The MCS-based allocation assumes that the VoIP packet size is constant and hence the MCS level is sufficient for indicating the allocation size. The authors propose to reduce the connection identifier size by allocation according to the MCS level rather by specific allocation size and location. The main drawback here is the assumption of a constant packet size, which may not fit a scenario where multiple coders are used, or even a single vocoder with multiple phases (e.g., the adaptive multi-rate AMR vocoder, see [24]). In addition the reduction is not substantial since a specific allocation is defined by

offset and length where the difference is minor. On the other hand, the proposed reduced MAP IE is still transmitted to the users, and hence generates unnecessary overhead compared to persistent schemes. In [22], the authors describe a persistent allocation messaging mechanism supported by WiMAX that was introduced to 802.16-2009 in [6]. We will further elaborate on the persistent allocation mechanism in 802.16 later on. In [20], the authors present a methodology to evaluate the VoIP capacity and performance of OFDMA systems by combining the queuing and interference analysis. The authors characterize the performance trade-offs of VoIP in OFDMA networks but do not propose a scheme of actual VoIP mapping. In [21], the authors present a persistent allocation approach in which an allocation is automatically made for both DL and for UL with some flexibility to enable statistical multiplexing. However, no particular information is presented on how selection of resources to users is given or how holes in the resources (2D) table are treated. In addition, there is no apparent treatment for an adaptive channel scenario. In [22], the authors present a persistent allocation scheme with MCS selection, specifically considering hybrid automatic repeat request (HARQ) retransmission mechanism that satisfies the given Quality of Service QoS requirements of the system. The authors acknowledge that MCS changes using instantaneous signal to interference plus noise ratio (SINR) is not a realistic approach for persistent allocation and propose employing MCS selection algorithms based on the assumption of the *Nakagami*-m SINR distribution and SINR averaging (where one algorithm approximates the other to achieve reduced complexity). This approach selects the best MCS for the session trying to maximize the rate while reducing the overhead of HARQ retransmission signaling. From all the reviewed solutions, [22] is the one closest to our approach for the decision of MCS selection. In [23] a slot-based persistent allocation scheme is presented. The authors point out that the current resource shifting and error-handling (defined below) mechanisms of 802.16 introduces additional controlling overhead. In the proposed scheme, a tree-like data structure is used to maintain the persistent allocation of slots to MSs and manage allocation of resource holes.

Persistent Allocation Support in IEEE 802.16 Systems

The 802.16 standard defines the option of dynamic scheduling. In this mode, the MAP message may point to sub-MAP messages, which are sent an MCS that is different from the most robust MCS used by the MAP message. These sub-MAP messages are targeted to MSs with matching CQI, which are able to interpret them. The sub-MAP usage was proposed in order to reduce the overhead of a single MAP message sent in the most robust — and hence resource expensive — mode. In addition, the reception of one sub-MAP is independent of the reception of other sub-MAPs messages. The dynamic scheduling mechanism was introduced for general efficiency purposes and does not necessary contribute much to the VoIP traffic model for two reasons. First, it supports up to three sub-MAPs at most and requires dynamic grouping of the MSs into the correct profiles. Second, for MSs with static channels, there is a minimal MAP overhead with a persistent allocation mechanism and hence using sub-MAPs does not necessary contribute to overhead reduction. In this work we do not use this option.

The IEEE 802.16 standard supports a specific mechanism of persistent allocation for optimizing the resource management of VoIP calls [6, 7]. As defined above, the persistent allocation in IEEE 802.16 takes advantage of the predictive nature of the VoIP traffic and allocates periodic resources to users with reduced signaling overhead. In addition, the standard provides means to efficiently eliminate resource holes using resource shifting and error handling. The error handling procedures are mainly used to handle sync losses between the BS and the MS, which may lead to resource collisions of different MSs if not handled properly. The provided error handling mechanisms include a MAP ACK channel for MSs to indicate the correct reception of allocation or de-allocation of a persistent resource, a MAP NACK channel for the MS to indicate the BS of problems in decoding MAP message and a change indicator for the MS to know if there is a risk of transmitting on a pre-allocated resource in case of MAP failure. Finally, the resource shifting procedure is used in case of de-allocation of persistent resources and shifting all the subsequent allocation over the released resource in a single MAP message. Fig. 1 [1] illustrates a resource hole and resource shifting.



■ ***Fig. 1.*** Illustration of resource hole and resource shifting for five resources

The goal of this research is to present a cross-layer persistent allocation solution, which fully complies with the messaging mechanism presented in [1, 2]. Our proposed algorithms solve the problem of how to map slots to MSs such that the overall overhead is decreased. We generalize the model given in [1] to support dynamic channels where the channel may change rapidly (i.e., mobile users in urban environment). This generalization enables us to model both codecs supporting silent suppression or adaptive rate and/or changes in the state of the mobile channel (and hence, supporting also variable MCS). More specifically, our proposed model enables us to uniformly represent an AMR codec with multiple states or multiple coders with different VoIP traffic parameters where none of the reviewed solution handles the multiple vocoder case. We show that our proposed algorithms reduce the mapping overhead for the dynamic cases as mentioned above, while keeping a simple and manageable resource allocations.

System Model

Notations and definitions

We consider an OFDMA UL mapper with a set of resources available for allocation. The resources are OFDMA sub-channels and time symbols where each combination of sub-channel and time symbol is called a slot. The slots are organized as a two-dimensional $M{\cdot}N$ table with M being the number of sub channels and N the number of OFDMA symbols. Unlike [25], for the problem at hand there is no need to use two-dimensional indices for the slots, thus each slot is assigned an index i, $i \in [0, M \cdot N - 1]$. The indices start with the lower left slot ($i = 0$) and the rows are scanned first from left to right and then from bottom up until the last slot with index $i = M \cdot N - 1$. A frame index is important for both the development of the mapper and the theoretical analysis, therefore the slots table in which the allocations take place is called an allocation table and is denoted by $A[k]$, where k is the frame index, $k \in \mathbb{N}$. A session r is defined as starting at a frame with index t_r to be characterized by a tuple $\langle N_r, d_r, p_r \rangle$, where $N_r \in \mathbb{N}$ is the number of bytes required for each transmission, d_r is the session duration expressed in terms of number of frames the session will last, and p_r is the frame period between two successive packets. The reason for defining N_r in terms of bytes rather than slots (as in [11]) is due to rate changes that depend on the channel state. A generalization of the above notation for mobile environments is as follows: a $\mathbb{N}$ session r starting at a frame with index t_r is characterized by a tuple $\langle N_r, d_r, p_r \rangle$ and a sequence $\langle N_r^1, d_r^1, p_r^1 \rangle$, $\langle N_r^2, d_r^2, p_r^2 \rangle$, ..., $\langle N_r^l, d_r^l, p_r^l \rangle$, where $\sum_{i=1}^{l} d_r^i = d_r$. This generalization enables us to model both codecs supporting silent suppression or adaptive rate and/or changes in the state of the mobile channel (hence, supporting also variable MCS). For codecs with silent suppression, the call duration is divided into "ON" periods with packet size N_{ON} with transmission period P_{ON} and to "OFF" (or silent) periods with packet size N_{OFF} with transmission period P_{OFF}. In each such session, we have $N_r^i \in \{N_{ON}, N_{OFF}\}$, and $p_r^i \in \{P_{ON}, P_{OFF}\}$, $\forall_i = 1, ..., k$. For mobile channels, where the MCS scheme may change due to a change in link quality we assume that each slot (i.e., the transmitted granularity unit) contains D bits. Assuming K MCS modes, each MCS scheme defines the number of bits per symbol m_j, $j = 1, ..., K$ according to the selected modulation and coding scheme. We assume, without loss of generality, that $m_1 \le m_2 \le ... \le m_K$, therefore, for MCS scheme j, a packet of size N_r^i will require $\bar{N}_r^i = \left\lceil \frac{N_r^i \cdot 8}{m_j \cdot D} \right\rceil$ slots.

The above generalization implies that each VoIP session r is represented as a sequence of l sub-sessions, where each sub-session i, $i = 1, ..., l$ is of duration d_r^i and requires $\bar{N}_r^i$ slots, depending on the vocoder's state and/or changes in channel conditions. From the point of view of the MOA that was presented in [11], such a change in a VoIP session was considered as mapping resources to a new call (signaling for a new call is not required) and the allocation of resources to the previous set of requirements is stopped. This approach is obviously sub-optimal since with each change in channel conditions a change in the MCS scheme is required, and therefore the BS needs to use additional system resources to manage such state changes. Next we discuss how to modify the MOA such that the channel variations are also considered.

For the sake of notation's simplicity, we ignore the sub-session indexing notation and treat a new subsession request r^i as a new session request r. We will use the specific sub-session indexing notation in an extended algorithm named EMOA, that will be described in full in part II of the paper. In EMOA the state of the previous sub-session is required for the algorithm decision process. We define a new session request (or a new request) for frame k as a session request (or just a request) r having $t_r = k$, and an active session request (or an active request) for frame k as a request a having $l = t_a + n \cdot p_a < t_a + d_a$ for some positive integer n.

An allocation instance for request r is a set of successive slots in $A[k]$, $S_r^k = \{s_i, ..., s_{i+\bar{N}_r - 1}\}$, satisfying $|S_r^k| \ge N_r$, for $k \in \{t_r, ..., t_r + d_r\}$ (i.e., the specific frames in which r is new or active); otherwise $S_r^k = \varnothing$.

An allocation instance S_r^k of request r in frame k will occupy slots having a contiguous set of indices. Therefore, the allocation can be referred to by the slot having the lowest indexed $s_i \in S_r^k$, and its size $\bar{N}_r$. An allocation for request r is a set of allocation instances, $S_r = \left\{S_r^{k_1}, S_r^{k_2}, ..., S_r^{k_q}\right\}$ for all relevant $A[k_i]$. An allocation set is a set of all allocation instances for some frame $A[k]$, $S^k = \left\{S_r^k\right\}$, where r is an active request at time k and $\left|S^k\right| = \sum_{S_r^k \in S^k} \left|S_r^k\right|$.

To each allocation table $A[k]$, there is an associated allocation map $M[k]$ containing all the UL-IEs that describe the transmissions of the MSs in the UL of frame k. An UL-IE for request r in frame k is denoted as $m_r^k \in M[k]$ and associated with S_r^k, where $m_r^k = \left\{s_i, \bar{N}_r\right\}$. The size of the UL-MAP of frame k is therefore $C_{map} + \sum_{S_r^k \neq \varnothing} \left|m_r^k\right|$, where C_{map} is some constant overhead that is independent of the number of IEs. The size of the UL-MAP is referred to as the mapping overhead. The effective area $EA[k]$ of allocation set S^k is defined as $|S^k|$ plus all the resources unused by the mapping algorithm (and thus are empty) $EA[k] = |S^k| + |A[k] - S^k|$. Any successive set of unused slots within the effective area is called a hole.

The **definition** of the efficient VoIP mapping problem is:

Given a sequence of allocation tables $A[k]$, $k = 0, 1, ..., K$, and a set of requests $R = \{r_1, r_2, ..., r_q\}$, find a set of non-overlapped allocations $S = \{S_1, S_2, ..., S_q\}$, (i.e., $S_i^k \bigcap S_j^k = \varnothing$, $\forall i \neq j$), such that the mapping overhead is minimized.

Traditional packing problems relate to an efficient assignment of multiple elements of some sort in a given container. In [11] we have solved a packing problem in consecutive time epochs, where the inputs in different time epochs are related. The main interest was both minimizing the effective area and the representation of the solution, which is the mapping overhead. In this paper, we present algorithms to solve the problem defined above with the mobility related extensions defined above. The dynamic nature of the channel needs to be taken into consideration by any future modification to the MOA. The main idea is to track the channel behavior of the MS and provide a predictor of the channel state changes probabilities and rates, which then can be used by the mapping algorithm. Such measures can indicate the stability of the channel, which can be used as an allocation grouping criteria to reduce the number of holes in the allocation table (see the entropy mapper that is presented in part II of the paper). In addition, we use the predictor of channel behavior, since rate assignments are later related to channel states. This also has implication on how to quantify the cost of state change in term of system resources and hence decides whether it is cost effective to perform an MCS state transition or just ignore it (see the EMOA mapper that will be represented in part II of the paper).

Mobile Channel Characterization

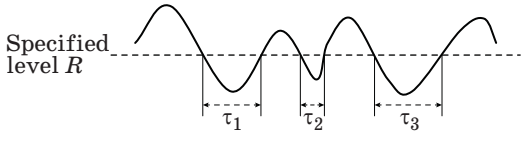
Markov chain modeling is a commonly used approach to study the behavior of fading signals in a mobile channel (see [26] for a survey on a finite state Markov chain for fading channels). It has been pointed out in [27] that a two-state binary Gilbert — Elliot Markov channel model [28, 29] is unable to capture dramatic changes in channel quality. Hence, we aim to capture the parameters of a fading wireless channel by using a FSMC defined over n states. In such a model, the fading process is partitioned into multiple discrete states, so that the dynamic behavior is captured by the transitions among the states. In contrast to physical models for mobile channels, FSMC is relatively simple, and has been widely adopted as a model for describing slow fading channels [27, 30]. This simplicity allows the practical implementation of channel state estimation for multiple MSs by maintaining a probabilities table for each MS (see sub-section "Extended-MOA" for further implementation details).

Finite-state Markov model of Rayleigh fading channels

To model the behavior of a wireless channel using a FSMC, the channel state transition probabilities must be associated with the chain. We do this by observing the effect the of level crossing rate (LCR) and average fade duration (AFD) on the signal-to-noise ratio (SNR) [31]. AFD and LCR describe the quality of the transmitted signal along the transmission channel. LCR is the level crossing rate of a signal envelope at a level R (Fig. 2) and is defined as the expected number of signal crossings of a level R in a positive going direction [32]:

$$LCR(R) = N_R = \int_0^\infty \dot{r} \cdot P(R, \dot{r}) \mathrm{d}\dot{r}, \qquad (1)$$

where $P(R, \dot{r})$ is the joint probability density function of the signal's envelope r and its time derivative $\dot{r}$ at $r = R$.



■ ***Fig. 2.*** Illustration of the definitions of LCR and AFD

As can be seen in Fig. 2, the AFD is the mean period of time for which the received signal envelope is below a specific level R. The AFD relates to the LCR according to:

$$AFD \equiv T_R = \frac{CDF(R)}{N_R}. \quad (2)$$

Here $CDF(R)$ is defined as the probability that the envelope of the received signal $r(t)$ does not exceed a specific level R, that is:

$$CDF(R) \equiv P_r(r \le R) = \frac{1}{T}\sum_{i=1}^{n}\tau_i, \quad (3)$$

where τ_i is the duration of the i'th fade (see Fig. 2) and T is the observation interval of the fading signal [32].

In a typical multi-path propagation environment the received signal envelope is Rayleigh distributed. For additive Gaussian noise the instantaneous received SNR γ is exponentially distributed with probability density function [31]:

$$p(\gamma) = \frac{1}{\gamma_0} e^{-\frac{\gamma}{\gamma_0}}, \quad (4)$$

where γ_0 is the average SNR; that is, $\gamma_0 \triangleq E[\gamma]$. The LCR for a Rayleigh fading channel for a given threshold Γ is defined by (see [31]):

$$N_\Gamma = \sqrt{\frac{2\pi\Gamma}{\gamma_0}} f_m e^{-\frac{\Gamma}{\gamma_0}},$$

where $f_m = f_0 v/w$ is the Doppler frequency shift, w is the propagation speed of the electromagnetic wave, v is the speed of the mobile user and f_0 is the carrier frequency.

The FSMC model can be built so as to represent the time-varying behavior of the channel by partitioning the received SNR into $K + 1$ intervals (states). If the fading process is slow enough, it is reasonable to assume that transitions will only occur between two adjacent states. This assumption was used in [27, 30]. Let $\Gamma_0 = 0 < \Gamma_1 < \Gamma_2 < ... < \Gamma_K < \Gamma_K = \infty$ be, the $K + 2$ thresholds that define the required partitioning. The channel is considered as "being in state k" if the instantaneous received SNR is between r_k and r_{k+1}. The states are ordered with decreasing average bit error rate (BER) values. We can calculate the steady state probabilities of the system using LCR as in [30]:

$$\pi_k = \int_{\Gamma_k}^{\Gamma_{k+1}} p(\gamma)\mathrm{d}\gamma = e^{-\frac{\Gamma_k}{\gamma_0}} - e^{-\frac{\Gamma_{k+1}}{\gamma_0}}, \quad (5)$$

where π_k is the steady state probability of the system being in state k. The average state duration $\overline{\tau}_k$ of the system in segment k is defined as

$$\overline{\tau}_k = \frac{\pi_k}{N_{\Gamma_k} + N_{\Gamma_{k+1}}} \quad (6)$$

and the transition probabilities can be estimated by:

$$p_{k,k+1} \approx \frac{N_{\Gamma_{k+1}} \cdot \tau}{\pi_k}, \quad k = 1, 2, ..., K; \quad (7)$$

$$p_{k,k-1} \approx \frac{N_{\Gamma_k} \cdot \tau}{\pi_k}, \quad k = 2, 3, ..., K + 1, \quad (8)$$

where τ is the system's sampling interval. We note that in the observed interval, the system may stay in the current state, hence

$$p_{k,k} = 1 - p_{k,k+1} - p_{k,k-1}. \quad (9)$$

Unlike [27, 30], we assume that the SNR interval sizes are given in advance as system parameters and are not a performance optimization goal, such as calibrating all states duration to have an equal value (see, e.g., [30]). This assumption is taken here to ease the implementation in practical systems in which supporting many MSs requires rate assignments in real-time.

Finite-state Markov model for general channels

The equations for both state and transitions probabilities can be generalized to model a general channel (i.e., not necessarily Rayleigh) as follows. Let us consider a random process $p(t)$, which has been partitioned into $K + 1$ discrete states as defined above. We denote the transition probability that the process $p(t)$ transits from state i to state j during the time displacement τ as $p_{i,j}$. Then, we denote the probability of $p(t)$ to be in state k at time t as $Pr(p \in [\Gamma_k, \Gamma_{k+1}))$. The transition probability can therefore be written as

$$p_{k,k+1} \approx \frac{LCR(\Gamma_{k+1}) \cdot \tau}{Pr\left(p \in [\Gamma_k, \Gamma_{k+1})\right)}, \quad k = 1, 2, ..., K; \quad (10)$$

$$p_{k,k-1} \approx \frac{LCR(\Gamma_k) \cdot \tau}{Pr\left(p \in [\Gamma_k, \Gamma_{k+1})\right)}, \quad k = 2, 3, ..., K + 1. \quad (11)$$

These expressions, according to equation (1), require the knowledge of the probability density function of $p(t)$. A similar argument holds for the steady-state probabilities.

Since the channel may not be easily expressed by an analytical formulation, we use channel SNR measurements to obtain an approximation to the related FSMC parameters. For ease of comprehension, we mention here some assumptions relevant to our model. First, we assume that the number of states and relevant thresholds are not configurable, thus given in advance as system related parameters. The reason behind this assumption is a common practice where system designers operate a set of offline simulations on various channel conditions with channel mixture scenarios. The results of such experiments are used to determine

a set of related adaptive MCS selections as a function of SNR (or BER) thresholds. Second, we assume that each state in our FSMC corresponds to a specific MCS scheme hence, given an instantaneous SNR, the FSMC state can indicate the appropriate rate (MCS scheme) to use. Last, we assume that the channel is stationary, which means that the state probabilities do not change through time or change slowly (i.e., the channel is quasi-stationary). Hence, the transition probabilities are not time-dependent. Due to the above assumptions and since we would like the implementation to be simple without the need for adjustments of the FSMC for some optimization criteria, such as maximizing system capacity, we use the following sampled based approximation.

Let T_s denote the symbol time, which means that the instantaneous SNR is measured for each symbol. Also, let $N_s(T) = T/T_s$ denote the number of samples measured during a time interval T. Last, let $N_s^k(T)$ denote the number of SNR measurements $\gamma \in [r_k, r_{k+1})$ during the measured interval T. Than the probability of state k is approximated by

$$\pi_k(T) \approx \frac{N_s^k(T)}{N_s(T)}. \qquad (12)$$

$N_{i\to j}(T)$ denotes the number of transitions from state i to state j during the measured interval T. Then we can define:

$$p_{i,j}(T) \approx \frac{N_{i\to j}(T)}{\sum_k N_{i\to k}(T)}. \qquad (13)$$

We note that as mentioned above, we limit the state transitions to adjacent states only, hence in the numerator of (13) $j = k + 1$ or $j = k - 1$. An additional argument, which is of interest, is the average return time to a state, since leaving that state. This problem is known as hitting time in general Markov chains theory [33]. Assume a given Markov chain with a finite state space $S = \{S_m | m \geq 0\}$ and a transition probability matrix **P**. For a subset $A \subseteq S$ the random variable $T_i(A) = \min\{j \geq 0 | S_0 = i, S_j \in A\}$ denotes the first hitting time of a state in A, starting from some initial state $S_i \in S$. It is easy to see that the expected hitting time can be determined by solving the system:

$$E_i\left(T(A)\right) = \begin{cases} 1 + \sum_j p_{i,j} E_j\left(T(A)\right), & i \notin A \\ 0, & i \in A \end{cases}, \qquad (14)$$

where $E_i(\cdot)$ is the expectation conditioned on $S_0 = i$. This means that in a general system with K states, by setting $A = S_0$ we will have to solve a system of K equations. In [33], some bounds on the hitting times are presented, but these are not practical for our purposes. Let us assume that $A = \{S_k\}$ and a bound on $E_i(T(A)) \leq \Theta$, that is, we are only interested in the cases where the expected hitting time of state S_k is lower than Θ. In addition, let us assume a Markov chain in which a transition is allowed only between two adjacent states. Under these assumptions, (14) can be modified to

$$E_{i,k} = \begin{cases} 1 + \sum_{j, |j-k| \leq \Theta} p_{i,j} E_{j,k}, & i \notin A \\ 0, & i \in A \end{cases} \qquad (15)$$

meaning that we need only to examine states with distance limit from state S_k. If the initial state S_i will always be either S_{k+1} or S_{k-1}, we can see that (15) will require to solve $\min\{\lfloor \Theta/2 \rfloor, m-k\}$ equations for the case of starting point S_{k+1}, and $\min\{\lfloor \Theta/2 \rfloor, k\}$ equations for the case of starting point S_{k-1}.

Conclusions

This paper discusses the mapping overhead problem in IEEE 802.16 OFDMA-based systems in the presence of mobile users. We present the problem of mapping overhead for VoIP sessions and show that it needs to be solved for practical systems such that the advantages of the OFDMA technology can be applied in future broadband wireless communication systems, both fixed and mobile. We are using the notion of "semi-fixed allocations" (which is similar to the persistent allocation) that enables the BS to discard IEs from the UL-MAP, thus reducing mapping overhead. Mobile users are mainly characterized by frequent MCS changes due to changes in channel states during the VoIP call. We provided a channel tracking and prediction model by employing the FSMC model and simple approximations of state probability and state transition probability. Therefore decision taken for a specific user needs to be adapted to changes in channel state. This development is a required building block towards a cross-layer oriented solution, which tracks the channel state in order to predict the expected cost of channel change on the overall system cost. One of the main observations is that in some cases, the BS should under-react to changes in channel conditions since it is more rewarding from a system-wise perspective.

References

1. **802.16-2009** IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE 802.16-2004, May 2009.
2. **IEEE Standard** for Local And Metropolitan Area Networks. Part 16: Air Interface for Broadband Wireless Access Systems. Amendment 3: Advanced Air Interface. IEEE 802.16m-2011, May 2011.

3. **Tech.** Spec. Group Radio Access Network; Evolved Universal Terrestrial Radio Access (Eutra); LTE Physical Layer General Description (Release 8). 3GPP TS 36.201 V8.1.0, November 2007.
4. **Ultra** Mobile Broadband (UMB) Air Interface Specification. 3GPP2 C.S0084 V2.0, September 2007.
5. **So J. W.** Performance Analysis of VoIP Services in the IEEE 802.16e OFDMA System with Inband Signaling. *IEEE Transactions on Vehicular Technology*, 2008, vol. 57, no. 3, pp. 1876–1886.
6. **Lu J., McBeath S., Barber P., Bourlas Y.,** et al. Persistent Allocation. IEEE C802.16maint-08/056r3, 2008.
7. **Fong M. H., Novak R., Mcbeath S., and Srinivasan R.** Improved VoIP Capacity in Mobile WiMAX Systems Using Persistent Resource Allocation. *IEEE Communications Magazine*, 2008, vol. 46, pp. 50–57.
8. **802.16-2004** IEEE Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE 802.16-2004, October 2004.
9. **IEEE** Standard for Local and Metropolitan Area Networks. Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation In Licensed Bands and Corrigendum 1. IEEE 802.16e-2005, February 2006.
10. **Das S. S., Ghosh P., and Chandhar P.** Estimation of Effective Radio Resource Usage for VoIP Scheduling in OFDMA Cellular Networks. *IEEE 75th Vehicular Technology Conf.* (VTC Spring), 2012, pp. 1–6.
11. **Kitroser I., Chai E., and Ben-Shimol Y.** Efficient Mapping of Multiple VoIP Vocoders in WiMAX Systems. *Wireless Communications and Mobile Computing*, 2009, vol. 11, no. 6, pp. 667–678.
12. **So J.** Scheduling and Capacity of VoIP Services (chapter) in Wireless OFDMA Systems. *VoIP Technologies*, InTech, 2011.
13. **Holma H., Kallio J., Kuusela M., Lundén P.,** et al. Voice Over IP (VoIP). In: *LTE for UMTS*. John Wiley & Sons, 2009, pp. 259–281.
14. **Jiang D., Wang H., Malkamaki E., and Tuomaala E.** Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE Systems. *Intern. Conf. on Wireless Communications, Networking and Mobile Computing*, September 2007, pp. 2861–2864.
15. **Saha S. and Quazi R.** Priority-Coupling-A Semi-Persistent Mac Scheduling Scheme for VoIP Traffic on 3G LTE. *10th Intern. Conf. on Telecommunications* (ConTEL 2009), June 2009, pp. 325–329.
16. **Konishi S., Komine T., and Shinbo H.** A Study on Persistent Scheduling for VoIP Services in UMB System. *Personal Indoor and Mobile Radio Communications* (PIMRC), 2008, pp. 1–5.
17. **McBeath S.,** at al. Efficient Bitmap Signaling for Voip in OFDMA. *Vehicular Technology Conf.*, October 2007, pp. 1867–1871.
18. **Shrivastava S. and Vannithamby R.** Group Scheduling for Improving VoIP Capacity. *IEEE 802.16e Networks, Vehicular Technology Conf.*, 2009, pp. 26–29.
19. **So J. W.** An Efficient Uplink Mapping Scheme for VoIP Services in the IEEE 802.16e OFDMA System. *International Journal of Electronics and Communications* (AEU), 2008, vol. 62, no. 10, pp. 768–776.
20. **Bi Q.,** et al. Performance and Capacity of Cellular OFDMA Systems with Voice-Over-IP Traffic. *IEEE Transactions on Vehicular Technology*, 2008, vol. 57, no. 6, pp. 3641–3652.
21. **Sambale K. and Klagges K.** Increasing the VoIP Capacity of WiMAX Systems Through Persistent Resource Allocation. *15th European Wireless Conf.*, 2009, pp. 308–313.
22. **Cho C., Jin H., Song N. O., and Sung D. K.** MCS Selection Algorithms for a Persistent Allocation Scheme to Accommodate VoIP Services in IEEE 802.16e OFDMA System. *Proc. of the IEEE 20th Intern. Symp. on Personal, Indoor and Mobile Radio Communications* (PIMRC), 2009, pp. 2122–2126.
23. **Tao M.-H. and Hsiao Y.-C.** Slot-Based Persistent Allocation for WiMAX Systems. *IEEE 21st Intern. Symp. on Personal Indoor and Mobile Radio Communications* (PIMRC), 2010, pp. 1271–1275.
24. **Speech** Codec Speech Processing Functions; Adaptive Multi-Rate (AMR) Speech Codec; Transcoding Functions. 3GPP TS 26.090 V10.0.0, March 2011.
25. **Ben-Shimol Y., Kitroser I., and Dinitz Y.** Two-Dimensional Mapping for Wireless OFDMA Systems. *IEEE Transactions on Broadcasting*, 2006, vol. 52, pp. 388–396.
26. **Sadeghi P., Kennedy R. A., Rapajic P. B., and Shams R.** Finite-State Markov Modeling of Fading Channels — A Survey of Principles and Applications. *IEEE Signal Processing Magazine*, 2008, vol. 25, no. 5, pp. 57–80.
27. **Wang H. and Moayeri N.** Finite-State Markov Channel — A Useful Model for Radio Communications Channels. *IEEE Transactions on Vehicular Technology*, 1995, vol. 44, pp. 163–171.
28. **Gilbert E. N.** Capacity of a Burst-Noise Channel. *Bell Syst. Tech. J.*, 1960, vol. 39, pp. 1253–1265.
29. **Elliott E. O.** Estimates of Error Rates for Codes on Burst-Noise Channels. *Bell Syst. Tech. J.*, 1963, vol. 42, pp. 1977–1997.
30. **Zhang Q. and Kassam S. A.** Finite-State Markov Model for Rayleigh Fading Channels. *IEEE Transactions on Communications,* 1999, vol. 47, no. 11, pp. 1688–1692.
31. **Jakes W. C. (ed).** *Microwave Mobile Communications*. 2nd ed. Wiley-IEEE Press, 1994. 656 p.
32. **Rappaport T. S.** *Wireless Communications Principles and Practice*. New York, NY, Prentice Hall, 1996. 641 p.
33. **Rego V.** Naive Asymptotics for Hitting Time Bounds. *Acta Informatica*, 1992, vol. 29, pp. 579–594.