CROSS-LAYER OPTIMIZATION FOR A MULTI-USER MIMO AUDIO TRANSMISSION

Marko Hennhöfer and Martin Haardt

Ilmenau University of Technology, Communications Research Laboratory, P. O. Box 10 05 65, 98684 Ilmenau, Germany
Phone: +49 (3677) 69-2615, Fax: +49 (3677) 69-1195
{marko.hennhoefer, martin.haardt}@tu-ilmenau.de

ABSTRACT

In this contribution we investigate the cross-layer issues occurring in a multi-user MIMO scenario. The scenario is based on a geometric model, thus containing realistic correlations in space, time, and frequency. As typical service we assume an audio transmission, where we evaluate the transmission quality by using the PEAQ (Perceptual Evaluation of Audio Quality) parameter. To build the bridge from the audio quality to the physical layer we derive a mapping from the PEAQ parameter to operation points consisting of required data rates and maximum allowable frame error rates. The optimization goal is to assure that each user attains a certain quality while minimizing the costs, i.e., bandwidth or transmission power. The parameters to be optimized are the multiple access scheme (SDMA, TDMA), the MIMO transmission mode (beamforming on one or two eigenbeams with diversity combining or multiple data streams, i.e., spatial multiplexing) and the choice of the modulation and coding scheme.

1. INTRODUCTION

During the last decades, the behavior and properties of the individual layers have been studied extensively. For wired transmissions the performance is quite deterministic as the channel is fairly constant compared to a wireless channel. To increase the capacity of wireless links, MIMO (multiple input / multiple output) systems are used. The optimization of a MIMO system is a challenging task due to the channel variations in time, space, and frequency, as well as the variety of the parameters to be optimized while assuring some quality of service (QoS) figures. As an example we use an audio streaming application. A good description of audio coding and real-time streaming is found in [1]. Several articles about cross-layer optimization for wireless systems are available [2, 3, 4, 5] but mainly focus on protocol aspects. The first article which deals with the challenges of the MIMO channel is found in [6]. In contrast to traditional physical layer optimization, where usually the throughput is maximized or the transmission power minimized, we want to start from the user point of view. Therefore, the main goal is to fulfill the QoS requirements of the application. To assess the manifold of parameters and influences for the optimization, Figure 1 visualizes the system structure.

If it is possible for a certain scenario to find a set of parameters to fulfill the QoS (Quality of Service) requirements of the individual users, we can further try to minimize the “costs” in terms of bandwidth or transmission power. We use the required bandwidth as cost function, as we assume a sufficiently large bandwidth for even strongly coded audio signals in 4G wireless systems. In Section 2 an audio quality parameter and the MP3 encoding is described. Section 3 highlights the typically used protocols for real-time applications in terms of functionality and the overhead. In Section 4 the effects of retransmission on the bandwidth are derived, followed by Section 5 where we use known results to derive the frame error rates. Three different MIMO transmission strategies are described in Section 6. Another parameter to be chosen is the multiple access technique where we restrict the discussion to TDMA and SDMA in Section 7. Simulation results are presented in Section 8, and Section 9 concludes the paper.

2. APPLICATION

From the variety of applications deployed in modern wireless systems we focused on streaming audio. The mapping of quality indicators to cost functions is a more general idea and can also be applied to other services such as video, voice traffic, or file sharing, by considering the different QoS parameters.

2.1. PEAQ - a Quality Measure

The optimization goal is to provide the individual users with a high quality audio stream while minimizing the costs of the transmission, e.g., bandwidth or transmit power. One problem is the definition of “high quality” for an audio transmission. In order to find an objective measure, we use the PEAQ (Perceptual Evaluation of Audio Quality) parameter as defined in the ITU-R recommendation [7]. This standard uses a number of psycho-acoustical measures which are combined to give a measure of the quality difference between two versions of a signal (a reference and a test signal). A detailed explanation of PEAQ is be-
Beyond the scope of this paper an can be found in [8]. The author also provides a Matlab implementation of PEAQ. To analyze the loss of quality we have to start with the encoding process. Using a song from a CD which was sampled at 48 kHz and stored as a WAV file, we encode it to MP3 files with different rates, ranging from 32 kbps up to 320 kbps. To use the PEAQ tool, the MP3s have to be decoded to WAV files again and are then compared to the original WAV file from the CD. Finally, the PEAQ parameter provides a measure for the quality difference between the reference and the sample file, which can be interpreted as summarized in Table 1.

<table>
<thead>
<tr>
<th>Difference grade</th>
<th>Description of impairments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Imperceptible</td>
</tr>
<tr>
<td>-1</td>
<td>Perceptible but not annoying</td>
</tr>
<tr>
<td>-2</td>
<td>Slightly annoying</td>
</tr>
<tr>
<td>-3</td>
<td>Annoying</td>
</tr>
<tr>
<td>-4</td>
<td>Very annoying</td>
</tr>
</tbody>
</table>

Table 1. Interpretation of the PEAQ parameter

A drawback of using PEAQ for the evaluation of the audio quality in practical (real-time) systems is the high computational complexity. Figure 2 shows the degradation due to the MP3 encoding and frame errors. A sample track of approximately 3 minutes length was used to extract the PEAQ difference grade.

Fig. 2. Audio quality degradation due to MP3 compression and frame errors.

Figure 2 shows the drastic decrease in quality if frames are lost. A lost frame results in a "clicking" sound which is indeed annoying. The performance will vary slightly depending on the encoder, whether the frame is just skipped, or if zeros or noise are inserted. From Figure 2 one would choose to transmit the 128 kbps stream and try to avoid frame losses at all by choosing a sufficiently long buffer. For example, if we assume an FER rate of $10^{-2}$ for the air interface and we want to stream our 3 minute sample file consisting of 7500 frames, we would have to retransmit 75 frames in order to achieve a frame error rate of 0 after the buffering stage. To guarantee a continuous stream we would have to buffer 75 frames in advance (worst case) which corresponds to 1.8 seconds. The buffer would then be "refilled" during the normal operation. Therefore some additional bandwidth has to be reserved.

2.2. Encoding

Before transmitting the audio data, it is encoded to the MP3 format. We used a freeware encoder to generate the MP3 frames. The used encoder produces MP3 frames of $T_F = 24$ ms duration, which results in 7500 MP3 frames for our 3 minute sample song. The number of information bits per MP3 frame, $N_F$, can then be calculated, depending on the rate $R = 1/T_{b,MP3}$, that was used for the encoding.

$$N_F = \frac{T_F}{T_{b,MP3}} = T_F R,$$

where $T_{b,MP3}$ is the bit duration after the encoding stage. To provide some numbers, the typical values are depicted in Table 2.

<table>
<thead>
<tr>
<th>MP3 rate, $R$</th>
<th>Frame size, $N_F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128 kbps</td>
<td>3072 bits</td>
</tr>
<tr>
<td>160 kbps</td>
<td>3940 bits</td>
</tr>
<tr>
<td>192 kbps</td>
<td>4608 bits</td>
</tr>
<tr>
<td>256 kbps</td>
<td>6144 bits</td>
</tr>
</tbody>
</table>

Table 2. Bits per MP3 frame ($T_F = 24\text{ms}$)

If the number of bits per packet is increased by introducing redundancy or protocol overhead, we still have to make sure that the corresponding MP3 data is transmitted within the time interval $T_F$. This is only possible by decreasing the bit duration, thus increasing the required bandwidth.

3. PROTOCOLS

In a first step we want to motivate the usage of the Real-Time Transport Protocol (RTP) in conjunction with the User Datagram Protocol (UDP) protocol. This protocol assumes that the Internet Protocol (IP) is used as the underlying protocol. The main reasons to choose RTP/UDP instead of the commonly used Transmission Control Protocol (TCP) are the lack of support for delay sensitive-services in TCP. The retransmission mechanisms of TCP cause undeterministic delays, whereas in UDP no retransmission functionalities are implemented, such that the more specific RTP can take care about the transmission control. Another advantage of UDP is the simplicity and therefore also the small header size of 8 bytes compared to 40 bytes for a TCP header. As an overview, the protocol stack is depicted in Figure 3.

Fig. 3. Protocol stack.

Examples how the MP3 data can be arranged in the RTP packets can be found in [4, 9]. Returning to our example we have to append the protocol overhead of 20, 8, and 12 bytes $(N_{ov} = 320 \text{ bit})$ for the IP, UDP and RTP protocol, respectively. The frame duration has to be maintained, as depicted in Figure 4. The bit duration after including the overhead $T_{b,ov}$ can then be

Fig. 4. Packet including overhead.
calculated.

\[ T_{b, re} = \frac{T_T}{N_{ov} + N_F(1 + \text{FER})} \]

In order to allow some retransmissions we also have to reserve some bits for the retransmission of data which will be described in the next section.

4. RETRANSMISSION

If we use the simplified assumption that a retransmitted frame is received correctly we can approximate the average overhead caused by retransmissions, \( N_{ov} \), by

\[ N_{ov} = \text{FER} \cdot N_F, \]

where FER is the average frame error rate. Therefore, the bit duration after considering the retransmission, \( T_{re} \), gets again decreased as depicted in Figure 4, what corresponds to an increased transmission bandwidth.

\[ T_{b, re} = \frac{T_T}{N_{ov} + N_F(1 + \text{FER})} \]

In order to map the overhead to a bandwidth expansion, it is necessary to introduce also coding and modulation briefly at this point. Assume a channel code with rate \( R_c \) and a modulation alphabet of size \( 2^k = M \), where \( k \) is the number of bits per symbol. For the coded bit duration and the symbol duration we get

\[ T_b = \frac{T_T R_c}{N_{ov} + N_F(1 + \text{FER})} \quad \text{with} \quad N_F = T_T R \]

and

\[ T_s = k T_b = \frac{k T_T R_c}{N_{ov} + T_T R(1 + \text{FER})}, \]

respectively. The required bandwidth, \( B \), is approximately \( 1/T_s \)

\[ B \approx \frac{k T_T R_c}{N_{ov} + T_T R(1 + \text{FER})} \]

With the constellation size, the code rate, the protocol overhead, the MP3 rate, and the mean FER we already have five parameters that influence our cost function, namely the required bandwidth. Some parameters, such as the protocol overhead or the mean FER can be treated as constant.

5. MODULATION AND CODING

Here we consider the AWGN (additive white Gaussian noise) case. The extension of the BER calculation to other channels is straightforward, e.g., by averaging over the received \( E_b/N_0 \) values. The uncoded symbol error probability, \( P_{s, unc} \), for \( M \)-ary QAM can be tightly upper bounded by [10]

\[ P_{s, unc} = 4Q \left( \frac{3k}{(M-1)N_0} E_b \right), \]

where \( k \) is the number of bits per symbol. Using the upper bound for the Q-function,

\[ Q(x) \leq \frac{1}{2} \exp \left( -\frac{x^2}{2} \right), \]

we can rewrite Equation (8) as

\[ P_{s, unc} \leq 2 \cdot \exp \left( -\frac{3k}{2(M-1)N_0} \frac{E_b}{N_0} \right). \]

Furthermore, we can include the coding with convolutional codes. To this end, we use the asymptotic coding gain for soft decision decoding, \( G_{a, soft, dB} \) [11]. The coding gain corresponds to a left shift of the BER curves in an \( E_b/N_0 \) plot.

\[ G_{a, soft, dB} = 10 \log_{10}(R_c d_t) \]

Here \( R_c \) and \( d_t \) are the code rate and the minimum free distance, respectively. The corresponding parameters of the commonly used codes are listed, for instance, in [10]. Therefore, the coded symbol error rate \( P_s \) results in

\[ P_s \leq \frac{2}{R_c d_t} \exp \left( -\frac{3k}{2(M-1)N_0} \frac{E_b}{N_0} \right). \]

As the number of symbols per frame is known, also the FER can be calculated.

\[ \text{FER} = 1 - \left( 1 - P_s \right)^{\frac{N_F}{T_T N_0}} \]

\[ = 1 - \left[ 1 - \frac{2}{R_c d_t} \exp \left( -\frac{3k}{2(M-1)N_0} \frac{E_b}{N_0} \right) \right]^{\frac{N_F}{T_T N_0}} \]

From the channel model we rather get SNR (signal-to-noise power ratio) values \( P/\sigma_n^2 \) instead of the bit energy per noise power spectral density \( E_b/N_0 \). Therefore, we use the following mapping

\[ P = \frac{E_b T_T}{T_{b, MP3}} N_i = E_b R N_i \]

\[ \sigma_n^2 = N_0 B = \frac{N_0}{k R_c} \left( N_{ov} + R(1 + \text{FER}) \right) \]

\[ P = \frac{E_k N_{ov} + RT_T}{N_0 N_{ov} + RT_T(1 + \text{FER})} \]

\[ \frac{E_k}{N_0} = \frac{P}{N_0 N_{ov} + RT_T(1 + \text{FER})} \]

where \( \eta \) is the SNR per transmit antenna and \( N_i \) corresponds to the number of transmit antennas. With Equations (13) and (17) the resulting frame error rate for a given link-SNR and modulation & coding scheme can be calculated. It is interesting to note that the instantaneous FER depends on the mean FER, as the mean FER causes an increased bandwidth due to retransmissions and therefore yields a higher noise variance.

6. MIMO MODES

From the great variety of MIMO transmission strategies we chose three different techniques, each of them aiming at different gains.

- **Antenna gain (ANT):** One data stream is transmitted along the strongest eigenmode, therefore sometimes also the expression dominant eigenmode transmission is used.
- **Spatial Diversity (DIV):** One data stream is transmitted along the two strongest eigenmodes.
- **Spatial Multiplexing (MUX):** Two data streams are transmitted along the two strongest eigenmodes.

In the following subsections the basic MIMO transmission model is expressed as

\[ y = \sqrt{\frac{P}{N_i}} H w s + n, \]

where \( y \) is the received vector of size \( N_i \times 1 \), \( P \) is the transmit power, \( w \) corresponds to a beamforming vector at the transmitter of size \( N_i \times 1 \), \( s \) is the symbol to be transmitted, and \( n \) is the spatially white ZMCSRG (zero-mean circularly symmetric complex Gaussian) noise vector.

6.1. Antenna Gain (ANT)

The received signal vector \( y \) is combined by a beamformer with weight vector \( g \). The beamformer output \( z = g^H y \) can then be expressed as

\[ z = \sqrt{\frac{P}{N_i}} g^H H w s + g^H n \]
To assess the SNR, we have to consider the powers of the signal and noise, respectively. With normalized symbols \( E\{s^*s\} = 1 \) and noise covariance \( E\{nn^H\} = \sigma_n^2 I_{N_t} \), the received power is

\[
E\{zz^*\} = \frac{P}{N_t} \left| g^H H w s \right|^2 + \sigma_n^2 |g|^2.
\] (20)

The SNR at the beamformer output, \( \eta_1 \), is then

\[
\eta_1 = \frac{P}{N_t \sigma_n^2} \left| g^H H w s \right|^2 / \sigma_n^2 \left| g \right|^2 .
\] (21)

The solution to the problem of finding SNR maximizing weights is provided by the singular value decomposition (SVD). The SVD is carried out for the channel matrix \( H \) (instantaneous CSI) or the spatial covariance matrices at the transmitter and receiver (long-term CSI), respectively. In case of the instantaneous CSI we can write the decomposition as

\[
U \Sigma V^H = H / N_t \times N_t.
\] (22)

Using the left and right singular vectors \( u \) and \( v \), corresponding to the largest singular value as beamforming weights, the input-output relationship simplifies to

\[
z = \sqrt{\frac{P}{N_t}} \sigma_s s + n , \quad E\{zz^*\} = \frac{P}{N_t} \sigma_s^2 + \sigma_n^2 \]

\[
\eta_1 = \frac{P}{N_t \sigma_n^2} \sigma_s^2 \] (23)

\[
\eta_2 = \frac{P}{N_t \sigma_n^2} \sigma_s^2 + \sigma_n^2 \] (24)

This expression is used now with Equation (17) to calculate the resulting \( E_b/N_0 \) for the detection process and to calculate the frame error rate.

\[ 6.2. \text{Spatial Diversity (DIV)} \]

Now the data stream is transmitted simultaneously on two eigenmodes, resulting in two links with an SNR of

\[
\eta_1 = \frac{P_1}{N_t \sigma_n^2} \sigma_s^1 \quad \text{and} \quad \eta_2 = \frac{P_2}{N_t \sigma_n^2} \sigma_s^2 \] (25)

Note that the transmit power has to be split up now onto both modes, \( P = P_1 + P_2 \). The optimum power allocation is provided by the water-filling solution, whereas in practical systems often an equal power distribution is used due to its low computational complexity. If we use \( P_1 = P_2 = P/2 \), the resulting SNR after a maximum ratio combiner is

\[
\eta_{\text{MRC}} = \frac{P}{2 N_t \sigma_n^2} (\sigma_s^2 + \sigma_n^2). \] (26)

Comparing (26) and (24), obviously (26) always results in a worse SNR. This is true, if the channel is known exactly. If long-term CSI is used to calculate the beamforming weights or channel estimation errors are considered, the probability that transmit energy is sent to a wrong mode of the actual channel is smaller for the spatial diversity case, as two modes are used and it is more unlikely that both modes are a bad match. This makes spatial diversity a good choice for fast time-varying, rich scattering channels.

\[ 6.3. \text{Spatial Multiplexing (MUX)} \]

In this case again two modes are used, but now different data streams are transmitted. To calculate the resulting FER, we can assume that the source data stream is split up into two substreams, while doubling the symbol duration, hence reducing the bandwidth by a factor of two. While in the spatial diversity case an error in one substream can be balanced by the other substream, we now loose the frame if an error occurs in either one of the substreams. Distributing the powers again evenly results in

\[
\eta_{1,2} = \frac{P}{2 N_t \sigma_n^2} \sigma_s^2 \quad \text{and} \quad \eta_{1,2} = \frac{P}{N_t \sigma_n^2} \sigma_s^2 .
\] (27)

Note that compared to Equation (25) the SNR per substream is improved due to the reduced bandwidth. The tradeoff is an increased symbol error probability

\[
P_e = P_{1,2} - P_{1,2} P_{1,2} \approx P_{1,2} + P_{2,2} .
\] (28)

where \( P_{1,2} \) and \( P_{2,2} \) are the symbol error probabilities for the individual substreams, resulting from the corresponding \( E_b/N_0 \) values.

\[ 7. \text{MULTIPLE ACCESS SCHEME} \]

The choice of the multiple access scheme is also influenced by a variety of parameters. For SDMA (space division multiple access) the number of users is crucial, as well as their spatial separation. For the simulations we used the block-diagonalization (BD) algorithm [12] which completely avoids interference between the users. It is assumed that the sum of all mobile terminal antenna elements is less or equal than the number of base station antennas. If this is not the case efficient grouping algorithms [13] can be used to sort the spatially well separated users into smaller groups and serve the groups, e.g., by TDMA (time division multiple access). In the presented example we used a uniform power allocation among the users. This is a sub optimal approach as we could actually also adapt the transmit powers jointly to achieve the QoS parameters for the individual users. As seen in Figure 6, user 2 can be served very well during the time he moves along his trajectory. The critical user is number 1 who uses strong coding while being far separated from the BS. A dynamic power allocation would yield a more balanced solution. In general TDMA has the drawback that it increases the bandwidth. For example, if two users should be served, it is necessary, to transmit the two MP3 frames within \( T_p \). This means we have to reduce the bit duration by a factor of two, which yields an increased bandwidth also by a factor of two. This means for the same transmission mode our cost function, the bandwidth, is already twice as high. Further on, as the transmit power is fixed we also get an SNR penalty from the doubling of the bandwidth, namely, an increased noise variance by a factor of two.

\[ 8. \text{SIMULATION RESULTS} \]

In this section we apply the derived equations to a specific scenario and determine the optimum operation modes. To model the scenario the geometry-based channel modelling tool, IImProp [14] was used. The parameters to be optimized are the multiple access scheme, whether to use TDMA or SDMA, as well as the modulation and coding scheme and the MIMO mode. The optimization is focused at a target frame error rate of \( 10^{-4} \) while minimizing the cost function, namely the required bandwidth. The transmit power, MP3 rate, and the overhead is considered as constant. To determine the eigenmodes, estimates of the long-term spatial covariance matrices at the transmitter and receiver where used, as this is a more realistic assumption as the perfect instantaneous CSI. Consider the following scenario,

\[ \text{Fig. 5. Simulated scenario.} \]

where the blue circles with the lines represent the receiving mobiles and their trajectories. Each mobile has a 4-element UCA (uniform circular array) and the base station is equipped with an 8-element UCA. User 1 travels along the x-axis with 90 km/h, whereas user 2 moves in y-direction with 15 km/h, which corresponds to 24 seconds for the whole scene. The red circle corresponds to the base station and the green clusters cause some
scattering. Using the time-varying (flat fading) MIMO channels generated by the model, we calculate the costs (bandwidth) for the different multiple access techniques as well as for numerous modulation and coding schemes.

Figure 6 shows, that SDMA always performs better than TDMA. The main reason for this is that in this configuration BD works really well, as the numbers of antenna elements is fairly high and the users are mostly well separated in space. Even around 8 seconds, where the users are almost at the same position, SDMA still outperforms TDMA. For TDMA the SNR penalty downgrades the performance, such that stronger codes have to be used to achieve the target FER which come along with higher costs (bandwidth). In reality SDMA would perform slightly worse, as in the used channel model double bounce reflections and dense multipath components are not modelled. For a certain point we are now able to determine which scheme achieves the target FER with the lowest costs. As the FER and the bandwidth where derived analytically, a simulation over a whole set of parameters can be done quite fast. To be more specific, we modified

- the MIMO mode: ANT, DIV, MUX,
- bits per symbol, k: 1, 2, ..., 8 (256-QAM)
- and the coderate, Rc (listed in [10]): 1 (uncoded), 1/2 (dc = 6), 1/3 (dc = 6), 1/4 (dc = 10), and 1/5 (dc = 13)

Figure 7 illustrates the used modes. The modes 1 to 8 are uncoded with constellation sizes from 2 up to 256, modes 9 to 16 use a rate 1/2 code and so on. For antenna gain (ANT) we add 100 to the mode, for DIV and MUX we add 200 and 300, respectively. E.g., mode 206 would be DIV, 64-QAM, uncoded. The numerous changes of the modes result from the temporal variation of the channel, and how well the eigenstructure of instantaneous channel fits to the one attained from the long-term spatial covariance matrices. In practice a mechanism to avoid such frequent changes would be implemented.

9. CONCLUSIONS

Starting from the user point of view, we determined the requirements for an acceptable audio stream. Using analytical results for the FER and the bandwidth, we are able to determine the optimum parameters for the transmission, namely the multiple access technique as well as the modulation and coding scheme. This is achieved by minimizing the bandwidth, which was used as cost function. Future work will include more detailed ARQ models as well as non-uniform power distributions among users while still maintaining the analytical description and mapping to the cost function.

10. REFERENCES