DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL 68 (2017), NO3, 167–179

VQ-based model for binary error process

Tibor Csóka, Jaroslav Polec, Filip Csóka, Kvetoslava Kotuliaková *

A variety of complex techniques, such as forward error correction (FEC), automatic repeat request (ARQ), hybrid ARQ or cross-layer optimization, require in their design and optimization phase a realistic model of binary error process present in a specific digital channel. Past and more recent modeling approaches focus on capturing one or more stochastic characteristics with precision sufficient for the desired model application, thereby applying concepts and methods severely limiting the model applicability (*eg* in the form of modeled process prerequisite expectations). The proposed novel concept utilizing a Vector Quantization (VQ)-based approach to binary process modeling offers a viable alternative capable of superior modeling of most commonly observed small- and large-scale stochastic characteristics of a binary error process on the digital channel. Precision of the proposed model was verified using multiple statistical distances against the data captured in a wireless sensor network logical channel trace. Furthermore, the Pearson's goodness of fit test of all model variants' output was performed to conclusively demonstrate usability of the model for realistic captured binary error process. Finally, the presented results prove the proposed model applicability and its ability to far surpass the capabilities of the reference Elliot's model.

K e y w o r d s: binary error; VQ, wireless channel, logical channel, error model, wireless sensor network

1 Introduction

Error characteristics of the binary channels, particularly the wireless channel, have been a focus of modeling ever since the error burst occurrence was proven to exhibit dependent behavior. Variety of different mathematical concepts and their combinations into more complex models have since been used to more or less precisely model different characteristics of the binary error process. The most widely accepted classification of error models is proposed by Kanal and Sastry in their review of channel error models [1] using a classification system based on the model's inner modeling principle: either generative (utilizing a generating "underlying mechanism") or descriptive (fit specific stochastic properties of the observed trace using empirical functions). More recent classifications have typically used the classification based on modeling employed mathematical concept to classify them as: pure (using only one mathematical method or a single principle) and extended (various model configurations primarily using cascading or modulating). The richness and variability of different types of stochastic behavior of the wireless channel leave open space for new models capable of surpassing the limitations of the current state-of-the-art.

2 Related work

Markov models were originally used to define the so called generative model group and have since become centric to multiple different model branches.The original proposals were based on the discrete time Markov chain, with the pioneering Gilbert's model [2]. Elliot suggested a modification [3] of Gilbert's error model by introducing the probability of generating an error also in the models good state. The following revolution came with Fritchman's model and particularly its simplification, the Simplified Fritchman's model, widely applied in high-frequency channel error modeling. More recent alternatives of Markov based models include the bipartite model [4], hierarchical Markov model [5] and extended models, such as cascaded Markov model [6] which employs parallel Gilbert's and Elliot's generators. Hidden Markov Models generate output trace using the same generative and mathematical principle, but the internal structure of the model is unknown and most approaches estimate its parameters using algorithms such as Baum-Welch or Turin-Sondhi [7]. A more recent addition to this group of models is the Double Embedded Processes based Hidden Markov Model [8]. Semi-Markov models were also promoted after [9] showed that packet loss can only be modeled using a time-inhomogeneous Markov chain.

Feasibility of Pareto distribution for error process modeling was successfully explored in a study by Ilyas and Radha [10] in their extensive research of errors on IEEE 802.15.4 LR-WPAN. Nogueira *et al* [11] offered a new perspective on empirical approach utilizing Markov concepts, a subgroup of Markov Arrival Processes (MAP) called Markov Modulated Poisson Process (MMPP) producing a hyper-exponentially distributed random variable. The problem of parameterizing MMPP models was addressed *eg* in [11–14].

Less commonly used types of models include the chaos-based models (*eg* [15–17] and a more complex approach

* Institute of Telecommunications, Slovak University of Technology in Bratislava, Ilkovičova 3, 841 04, Bratislava, Slovakia, csoka@ut.fei.stuba.sk; polec@ktl.elf.stuba.sk; filip.csoka@gmail.com; kkotul@ktl.elf.stuba.sk

DOI: 10.1515/jee-2017–0025, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X

by Kopke *et al* in [18]), the discrete process-based generative model (DPBGM) based on the principle of Rice's sum of sinusoids (extensively described in [19] and [20], both papers are validating the model proposed in [21] on a real EGPRS channel trace), Stochastic context free grammars (SCFG), fractal models, multi-fractal wavelet model [22] and improvements of existing models by new concepts, such as genetic algorithms (*eg* [23]).

3 Binary error process trace

An error process on a digital communication link can be considered a binary discrete-time stochastic process. If I is a countable set of integers $t \in I$, a_t the digital input sequence, b_t the corresponding output sequence and n_t the noise sequence also referred to as trace, then

$$b_t = a_t + n_t \,. \tag{1}$$

A correctly received bit is in a trace represented by "0" and an incorrectly received bit is represented by "1". Error modeling then becomes equivalent to statistically correct modeling of the trace characteristics.

Consecutive sequence of "1" is called an error burst. An error gap may be defined as a sequence of consecutive "0" between two "1" and represents the distance of two neighboring error bursts in bits. Empirically the shortest error gap or error burst has length 1 [1]. The error overflow assumption stating that the last "1" of the previous packet and the first "1"" in the following error packet are not part of the same burst error, is considered in this paper as well.

The trace used in this paper was captured in a WSN laboratory environment on a non-line-of-sight (NLOS) channel in indoor environment described in [24]. Verification of empirical and Elliot's model application on the same binary trace as the one used in this paper is presented in [25].

4 Proposed class of VQ-based models

The proposed model can be logically divided into 2 logical and functional parts with distinct roles within the model: randomness introduction and modeling transformation.

Randomness is inserted into the model by an arbitrary stochastic concept transforming an input from a Random Number Generator (RNG). The role of randomness introduction in the proposed model is performed by Markov chain, whose states each represent a group of vectors from the codebook. Transitions are established by observing the binary vectors in the trace and transitions of their corresponding abstracted states. Randomness introduction part of the model is thereby also responsible for capture and modeling of the large-scale stochastic behavior present in the observed trace. A variety of different stochastic concepts could be used instead of Markov chains, however, as demonstrated by the results of the models' output, Markov chains are sufficient for modeling the experimental WSN trace.

Small-scale stochastic behavior is modeled by the distribution of individual vectors within each state. Therefore, as such, not only the transitions of states representing classes within codebook must be observed, but also the individual stochastic behavior of vectors from each class within their respective classes.

The following chapters in this paper demonstrate the construction of two different types of codebooks, both of which share the same principle of modeling the small-scale error process by using the inverse generating method to find the most suitable vector to generate.

4.1 Vector quantizer model based on Hadamard codebook (HVQ)

The proposed HVQ model uses a well know VQ concept and introduces a novel approach to codebook construction for binary vector generation.

The codebook's vectors are constructed from an arbitrary basis that allows multiresolution by using vectors of different dimensions. For practical purposes, the codebook could be constructed from optimal vectors obtained from large number of measurements. Due to utilization of binary vectors for codebook construction, the lossy VQ compression method becomes lossless regarding the information content, provided that the binary vector set in the codebook forms a basis. Because the modeled binary trace can be entirely described by a codebook containing a finite number of vectors with different length, this model retains all information about the observed binary process.

A novel idea introduced for modeling arbitrary binary traces using a HVQ model is the codebook construction exploiting the properties of binary sets and maximizing both parameterization and generation computation resource efficiency. Assuming that the burst and gap processes do not necessarily have to be independent leads to formulation of a claim, that there are binary runs of higher than minimal order having increased probability of occurrence in the observed binary trace than other runs. It is preferable to construct such codebook that contains not only as small basis set of a chosen binary space, as possible, but also include the vectors representing the frequently occurring identified specific runs.

Such a set that contains basis vectors from a particular binary space and additional vectors from the same or other binary spaces exhibits overcompleteness. The significant advantage of implementing the proposed codebook lies in 2 important aspects: ability to represent all binary vectors of specific lengths with the basis vectors contained in the codebook and ability to represent specific longer binary sequences using the additional codebook vectors. Different approaches to basis set construction can therefore be chosen:

- Custom basis vector set

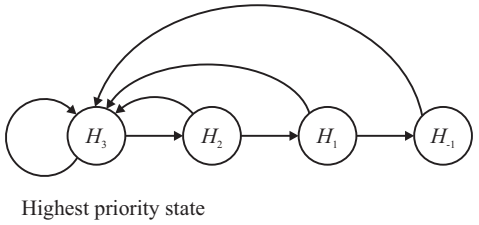

Fig. 1. Abstract DTMC for the codebook defined by BE

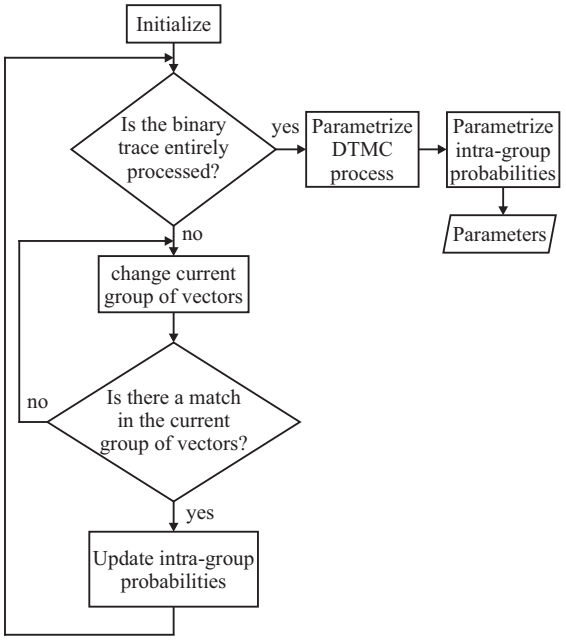

Fig. 2. Parameterization of the HVQ

- Standard basis vector set

HVQ uses vectors from Hadamard matrices of different orders, therefore falls into the second category.

Definition of a suitable codebook does not guarantee an efficient model. However, a combination of random process selecting vectors from the codebook in the generation process is enabled by the abstract DTMC, whose states represent different groups of vectors within the codebook. Particularly interesting is the case, in which the vectors are organized into groups based on their length or focus. Binary vector assignment in the parameterization phase is equivalent with transition to the first state of the multiresolution chain.

Within each state (group of vectors), vectors are assigned different generating probabilities based on their occurrence in trace, effectively producing an intrastate generating process. Stochastic process described by such a DTMC therefore represents a VQ modulated Markov process.

1) Example basis construction

The proof-of-concept demonstrated in this paper was realized using a codebook constructed from the standard Hadarmard matrix. The Hadamard matrix boasts many different beneficial properties, of which the most interesting ones for purposes of error burst and gap modeling is the well balanced ratio of binary values in the matrix and ability to capture both short and long runs of binary values. The Hadamard matrix of order 1 (note that the matrix numbering is different from Hadamard's original numbering) is

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{2}$$

The core of HVQ model's codebook is made of vectors contained in H_1 and $-H_1$, because together they form a basis and thus guarantee that every binary vector of length 2 is uniquely assigned to one of the codebook vectors.

The Hadamard matrix of order n is defined as

$$H_n = \begin{pmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{pmatrix}. \tag{3}$$

Adding suitable vectors from Hadamard matrices of higher order extends the codebook's multiresolution capability.

2) Parameterization procedure

Parameterization procedure (Fig. 2) starts in the initialization block representing the creation or copying of the Hadamard matrices into the codebook.

Three different HVQ codebooks were used to prove the concept; they are composed of the following vectors

$$\begin{aligned} BE &= \{H_3, H_2, H_1, -H_1\}, \\ BG &= \{-H_3, -H_2, -H_1, H_1\}, \\ D &= \{-H_3, H_3, -H_2, H_2, -H_1, H_1\}. \end{aligned}$$

Each of the vectors in the Hadamard matrix is represented by a state in the abstract Markov chain that introduces randomness into the model. The transitions among the states are strictly limited to transitions from the state with higher priority to the nearest state with lower priority, unless a precise match of the analyzed binary vector is identified, at which point the system returns to the state with the highest priority. Priority is assigned by the order of vector groups, with the leftmost element of each codebook having the highest priority and rightmost elements having the lowest priority. This principle is depicted for the vector codebook BE in Fig. 1.

The binary set is sequentially and iteratively compared on the binary vector basis to the vectors from each of the groups contained in the codebook. A match is followed with increasing the probability of the state corresponding to the group containing the vector and increasing individual probability of identified vector's generation within the group (state). If there is no match for the vectors of the current group, a lower priority group is selected and the sample from the trace compared to the new group vectors.

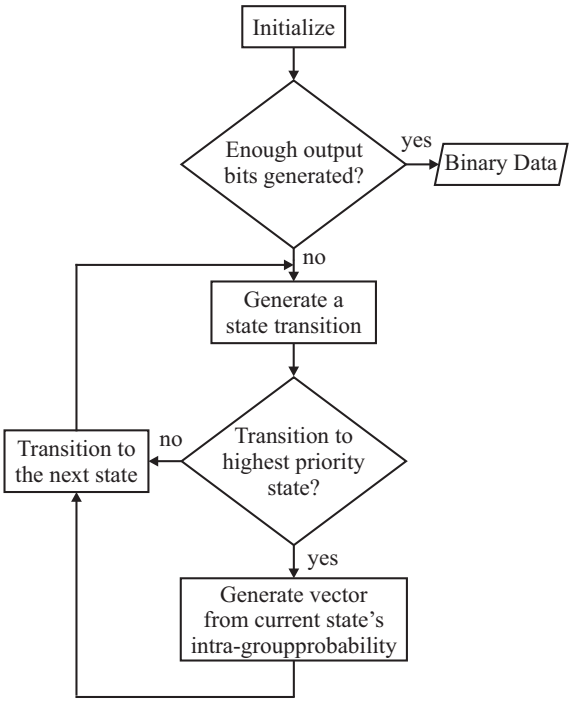

Fig. 3. HVQ generation procedure

Because the basis set is overcomplete, there is a match for every binary vector combination that can occur, if not in the higher order vectors, then in the lower order Hadamard matrices.

3) Generation procedure

In the initialization phase the entire Hadamard basis set is constructed and filled into the codebook (Fig. 3). Secondly, a random initial state of the abstracted modulated Markov chain is chosen.

The generation procedure introduces randomness via the Markov chain modulated by the VQ codebook and generation probabilities assigned to each individual vector within each group (state of the Markov chain) in the codebook.

The generation process is controlled by the Markovian transition matrix. Each iteration the RNG produces a value used by the current state (representing a group of codebook vectors) to establish its transition to the next state. If and only if this next transition is to the highest priority state, the group is used to generate a binary vector. It does so by using a new RNG value to produce a vector based on the intra-group (intra-state) vector probabilities established in the parameterization process. Regardless of whether the generation in the current iteration occurred or not, the state transition defined by the first RNG occurs at the end of the iteration.

The process is repeated until the desired number of bits are generated.

4.2 Proposed classification-based VQ model (CBVQM)

Classification-based binary VQ model represents a novel approach to modeling binary error process captured in form of an arbitrary binary trace (obtained using (1) from the binary transmitter output and receiver input) of any binary channel type. The channel, however, has to be stationary, due to the stationary nature of the proposed model itself. Adaptive extensions are possible, but are not focus of this paper.

Error burst and gap processes are considered in the general case, *ie* they both can be dependent and independent, meaning that generation of binary bursts and gaps is not separated, bursts and gaps are being generated together in the same iteration and instance as parts of a greater unit referred from now on as generated vector.

Classification is used to construct the codebook used in the generation process in such a way, that binary codebook vectors belong to at least one of the classes. In order to achieve maximal precision, the total number of different dominant stochastic sub-processes forming the error bursts and gaps should be less than or equal to the total number of desired classes. In order to retain overall stochastic behavior and individual burst and gap distributions present in the trace, higher model precision is achieved by capturing the histogram characteristic (relative number of runs) of the observed burst and gap processes, instead of the burst and gap run sequences. Captured histograms of binary vectors from the trace with fixed size are transformed into feature vectors (FV) necessary for the classification process.

Utilizing a classification approach allows great variability in applying different classification techniques (results present in this paper were obtained using kmeans) and sorting techniques (*eg* K nearest neighbors (KNN)).

As with every classification problem, one of the key issues of this model is establishing the optimal number of classes. Due to the nature of combined binary burst and gap processes, this is most reasonably performed by running classification process with different settings and choosing the results with the best fit for the particular channel.

The choice of distance metric that can be used to group various types of distance rules of different fixed-size vectors in the trace is a factor affecting the precision of the resulting model as well. Assuming that part of the process invariance is removed by employing histogram representation instead of occurrence order in FV construction, Euclidean metric is considered a sufficient distance measure for classification purposes and is therefore also used in this paper.

Randomness is added to the CBVQM model by abstracting the identified classes as states of a discrete-time Markov chain (DTMC), thus limiting the class transitioning process to geometrical distribution. This can, of course, be improved by using a different stochastic concept for state transitions, but experiments have demonstrated, that application of Markov chain random process is sufficient for binary error channel modeling purposes.

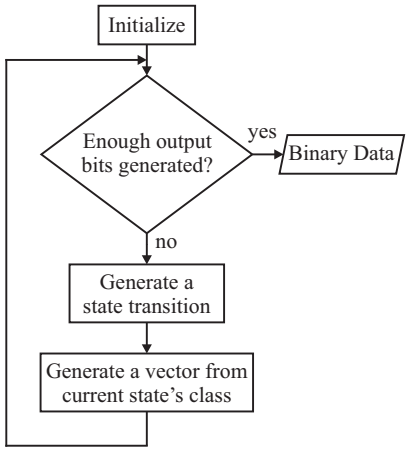

Fig. 4. Generation from the CBVQM model

4) Parameterization procedure

The CBVQM model parameterization process (Fig. 5) starts with input variable initialization. The binary trace with length n is partitioned into $\lfloor n/l \rfloor$ temporary binary vectors of length l, each representing a part of the trace with all characteristic features of the binary error process contained. Information about the binary process' stochastic characteristics that cannot be contained within the binary sequences with length l and can only be observed on binary sequences longer than the length of the temporary binary vector, will not be captured by the proposed model's parameterization process and therefore can only occur as random behavior in the generated data.

Temporary binary vectors can be used as FV in the classification process; however, the order in which the bursts and gaps appear is time-variant. Therefore, the same burst and gap binary sequence shifted by one bit would produce a completely different FV. This is an undesirable effect, because such FV obviously does not retain any information about stochastic distributions of error burst and gaps within the temporary binary vectors; rather it stores information about their order. In order to retain as much information about the stochastic behavior contained within each of the temporary binary vectors, as possible, the FV are not constructed from the direct sequence of bursts and gaps, but instead from its histogram representations, one for the bursts and one for the gaps. At this point a compression factor could be introduced, *eg* histograms would be shorter than the length of the temporary binary vectors, thereby assigning any occurrence of value larger than the last bin's size to the last bin.

FV is thus obtained as transformation of the temporary binary vector into a histogram of bursts and histogram of gaps, which are in the FV ordered as: (burst histogram, gap histogram). The classification matrix necessary as an input for the chosen classification technique is constructed by placing each FV into a different row of the classification matrix.

The chosen classification method with the specified metric is then used to sort FVs into the selected number of classes using the specified metric. Each class is uniquely defined by its centroid value representation identified during the classification procedure, where the centroid, similarly as the feature vector, represents the duplet (centroid burst histogram, centroid gap histogram). It should be noted, that centroid histograms do not have to directly resemble the histograms in the FVs and can create unique histograms representing distributions different from each FV.

The temporary binary sequences are after a successful classification process replaced by the class to which each individual FV is the closest (distance-wise respective to the selected classification metric). Each class represents a state of the DTMC in the generation procedure, therefore the transition probabilities of the observed trace need to be established. Considering that the length of the binary

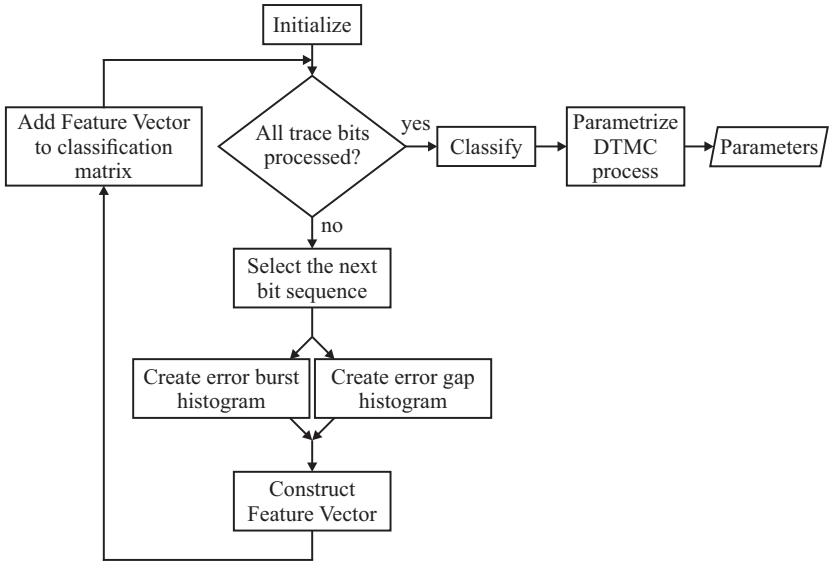

Fig. 5. Parameterization of the CBVQM model

trace is sufficiently long to capture the inter-class behavior, individual state transitions can be established directly from the class sequence obtained by temporary binary sequence substitution with class identifier. The result of the DTMC state transition parameterization is the transition probability matrix of the abstracted Markov process.

5) Generation procedure

The initialization within the generation process (Fig. 4) starts with the creation of an entire set of vectors from the binary set of order l. By assigning each of the vectors from the complete binary alphabet (all binary vectors of length l) into one of the classes defined by the centroid using the same metric, the model is capable of producing every binary sequence of the specified length. A classifier such as *KNN* can be used for this purpose.

The data set could however, due to the inclusion of all vectors from the complete alphabet for the specific vector length, exhibit high imprecision caused by not compensating to reflect the actual probability of generation for each of the vectors within a single class.

One of the relevant factors that could be used to generate the probability of generation for an individual vector is its distance to the centroid. Because all binary vectors of length l are represented in the binary alphabet, it is possible that a number of vectors is not within a reasonable distance to any of the centroid, yet the classification process will assign it to a particular group because of its closest proximity to it, in order to increase the generating probability of those vectors that are closer to the centroid and decrease the generating probability for vectors that are further away (distance penalization). Thus, the generating probability of the i-th vector in the j-th class depends on the second power of the inverse distance d_i of the vector from the centroid

$$p_i \approx \frac{1}{d_i^2}\,. \tag{4}$$

The distance proposed by this metric proved sufficient for binary error burst and gap modeling, but is subject to further discussion and exchange for a different metric that could prove to produce even more precise results.

All probabilities p_i for each state are used to produce a vector generating histogram that can be transformed into the CDF of the intra-state vector generating process.

A distinct complication arises, particularly, if the total number of classes is higher than the number of components of an error burst and gap process. In such a case, multiple centroids close to each other are identified in the trace, but the process of assigning the vectors from a complete set based only on the nearest neighbor would assign the binary vector to the nearest class. That would, however, undermine the distance concept of generating probability calculation. Two readily available solutions could rectify this problem: reduction of total number of classes or assignment of the same vector to multiple classes.

Multiple vector class assignment is a faster solution that retains the proposed number of classes, where any distance shorter than the nearest neighbor for all vector assignments must be considered, resulting in the possible presence of a single vector in multiple classes with different generating probabilities.

Lastly, an arbitrary state from the abstracted Markov chain representing the stochastic transitioning process is selected as the starting state, concluding the initialization phase of the generation process.

Once the initialization stage is finished and the generating set has been initialized and configured to reflect the binary data trace parameters, the generation process is relatively simple and can be summed up in two distinct steps. Firstly, a state transition based on the output of the RNG is produced. Then, after each state transition, the destination state's intrastate generating CDF (based on individual vector probabilities of occurrence p_i) is used to produce a binary vector from that state using the inverse method, repeating the process for as many bits as need to be generated.

5 Statistical distances for result analysis

Divergences were considered for analysis of the modeled results regarding their ability to capture non-complete histograms with limited ability to partition the bin space.

Following chosen statistical distances serve as reference values for establishing the quality and precision of the proposed models.

5.1 Hellinger distance (HD)

Hellinger distance is a special case of the β-divergence for $\beta - 1/2$ and is defined for discrete measurements as

$$D_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}. \tag{5}$$

5.2 Divergence χ^2

A special case of χ^a-divergence, where $a = 2$ is defined as

$$\chi^2(P,Q) = \sum_i \frac{(p_i - q_i)^2}{q_i}\,. \tag{6}$$

5.3 Jeffrey divergence (JD)

Jeffrey divergence is an improvement of Kullback-Leibler divergence which improved on its deficiency in evaluating histograms with various bin occupancies. JD does not take into account any bins that have zero occupancy in any of the compared histograms. This divergence is relatively underused despite its beneficiary properties and it was purposefully chosen for verifying the model applicability due to the unevenly generated data in histogram bins that cannot guarantee that the same bins

will be occupied in both the reference and the modeled histograms. Thus, JD avoids the problem of division with 0 present in KL-divergence. Jeffrey divergence is

$$D_J(P,Q) = \sum_i \Big(p_i \log\log \frac{q_i}{m_i} + q_i \log\log \frac{p_i}{m_i} \Big) \qquad (7)$$

where $m_i = (p_i + q_i)/2$.

5.4 Mean-squared error (MSE)

The MSE corresponds to the second moment of the error representing the variance of the compared model. Given histograms p and q, where p represents the reference observation, the MSE of the model q is

$$MSE(P,Q) = \frac{1}{n} \sum_i SE_i = \frac{1}{n} \sum_i (p_i - q_i)^2 \qquad (8)$$

where n is the total number of unique bins and SE is the squared error of the i-th bin observation.

5.5 Pearson χ^2 goodness of fit test

The null hypothesis of the Pearson χ^2 goodness of fit test is assumption that the observation originates from a theoretical probability distribution produced by a stationary ergodic source. The testing statistic is calculated as [26]

$$\chi^2 = \sum_{\forall i} \frac{(f_i - NP_i)^2}{NP_i} \qquad (9)$$

where f_i is the observed empirical frequency of values of the i-th class, P_i the theoretical probability of values in the i-th class of the total set divided into n classes and N is the size of the observed data set. Validity of the null hypothesis asymptotically approximates the χ^2 distribution with $n - p - 1$ degrees of freedom (p is the number of estimated parameters).

The χ^2 test is mathematically correct, however similarly as other statistical tests, the null hypothesis is rejected for large data sets, such as the one produced by the wireless channel model. This effect can however be mitigated by the coefficient of discrepancy C [26], applicable only under the assumption that the χ^2 statistic rises linearly with the number of elements N in the set, and the theoretical relative probabilities are stationary, then the discrepancy coefficient

$$C = \frac{\chi^2}{N}. \qquad (10)$$

According to [26], the acceptable values for the goodness of fit test are $C \leq 0.05$ for large data sets (such as the one produced by the proposed generators).

6 Results

The parameters for each model are not included (as they can be quite extensive), but can be obtained by the parameterization process described in Section 4.1.

Although the real analysis is performed on discrete data, for a better visual representation all stochastic processes were interpolated with the cubic spline function. Histograms are depicted with relative probability values instead of the absolute number of elements in the bin to better visualize the process' PDF.

6.1 HVQ

HVQ, unlike the empirical models and Elliot's model, successfully passes all goodness of fit tests in case of HVQ-BG and HVQ-D with HVQ-D being the most successful one according to the majority of the distance metrics. Considering its superior quality and error process capturing ability, the HVQ-D and HVQ-BG can both be considered a suitable and verified replacement for the reference Elliot's model.

1) Cluster error analysis

The cluster error analysis of data produced by three different variants of the HVQ can be seen in Fig. 6. The three variants produce significantly different results. Rather surprising is the low precision of the basic version with emphasis on the error burst resolution. The only conclusion that can explain why such a difference in cluster probability fitting precision between a version prioritizing higher resolution of errors (HVQ-BE) and versions prioritizing higher resolution of gaps (HVQ-BG, HVQ-D) occurred is because the gaps appear to be more important for creating the overall cluster probability. The total bit error rate is not very high in the observed channel, only 11.124 %, hence the incomparably bad performance could be explained by the fact the error bursts form only slightly more than the tenth of all binary symbols present in the trace. Therefore, by increasing the burst resolution at the expense of the binary error gap resolution, the overall cluster error probability characteristic suffers.

Quantification of the visual representation from the selected metrics (Tab. 1) proves the conclusions obtained from the visual observation of the cluster error probability. Classification based binary model version with the highest emphasis on multiresolution capture of error bursts produces the worst results. Although the Fig. 6. cannot be used to conclusively establish, which of the other two versions is superior, the calculated distances are in all cases in favor of the HVQ-D version, which employs multiple higher order matrices for multiresolution of both errors and gaps. The results produced in the cluster error probability analysis are even superior to those produced by the reference Elliot's model, hence regarding the binary cluster error probability, the HVQ versions BG and D are a suitable substitution for the reference model in the domain of cluster error generation, producing a better fit.

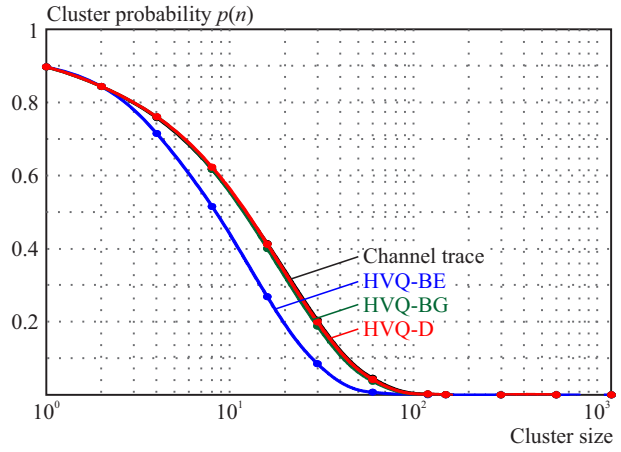

Fig. 6. Cluster error probability curves $p(n)$ for selected models

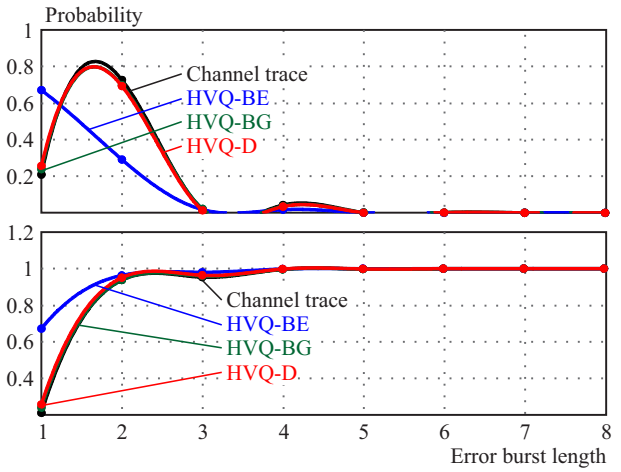

Fig. 7. Error burst histogram (up) and CDF (down) for all models and the real channel data

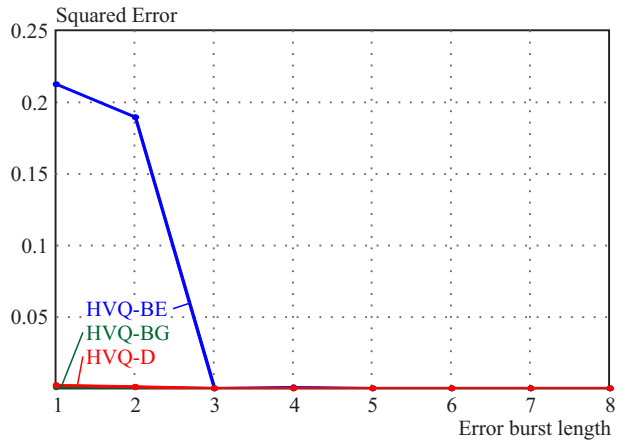

Fig. 8. Error burst SE curves for all models

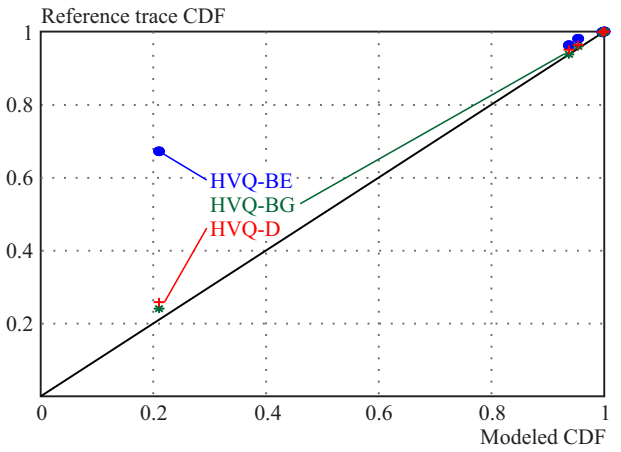

Fig. 9. P–P plot of the error burst models compared to the reference trace

Table 1. Distance metric evaluation of probability $p(n)$.

	JD (D_J)	HD (D_H)	MSE
HVQ-BE	0.02726	0.18014	0.04886
HVQ-BG	0.00044	0.02247	0.00051
HVQ-D	0.00004	0.00708	0.00008

2) Error burst analysis

The results of binary error burst process within the generated data also confirms high quality of the proposed HVQ. Apart from the low gap resolution version HVQ-BE, the other versions demonstrate excellent precision in capturing the nature of the burst error, as demonstrated in both the PDF and CDF of the error burst process (Fig. 7). Surprisingly enough, not even the higher resolution of the HVQ-BE model in error burst analysis was sufficient in producing a reliable characteristic.

Table 2. Distance metric evaluation of error burst and gap fit

	Error burst				Error gap			
	JD (D_J)	HD (D_H)	MSE	$C(\chi^2)$	JD (D_J)	HD (D_H)	MSE	$C(\chi^2)$
HVQ-BE	0.09965	0.34254	0.40329	1.29292	0.02529	0.17391	0.01088	0.01408
HVQ-BG	0.00093	0.03277	0.00174	0.00874	0.02616	0.17542	0.00788	0.00073
HVQ-D	0.00168	0.04441	0.00341	0.00335	0.02182	0.15993	0.00791	0.00031

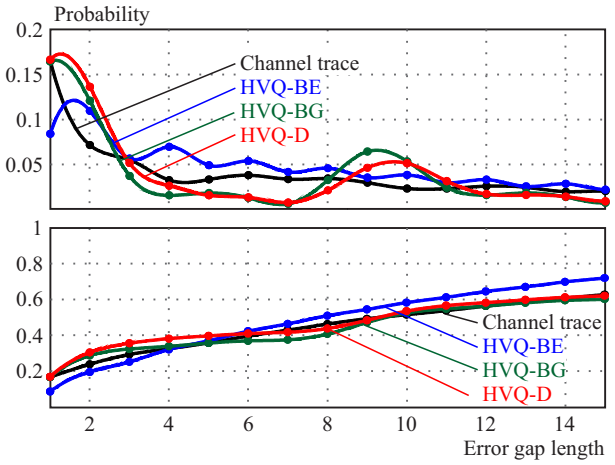

Fig. 10. Error gap histogram(up) and CDF (down) for all models and the real channel data

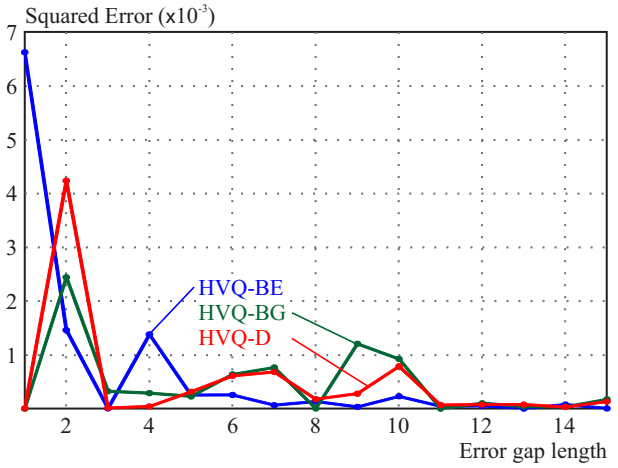

Fig. 11. Error gap SE curves for all models

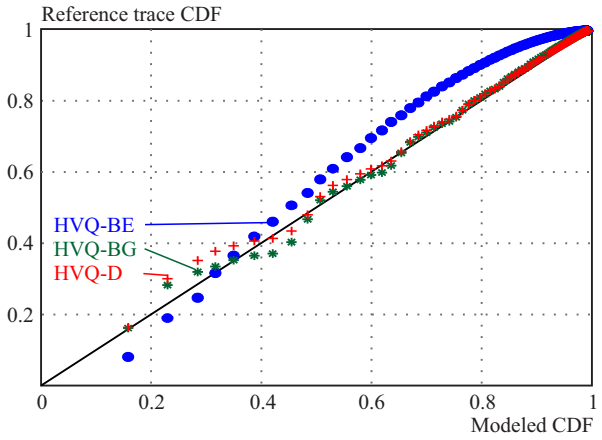

Fig. 12. P-P plot of the error gap models compared to the reference trace

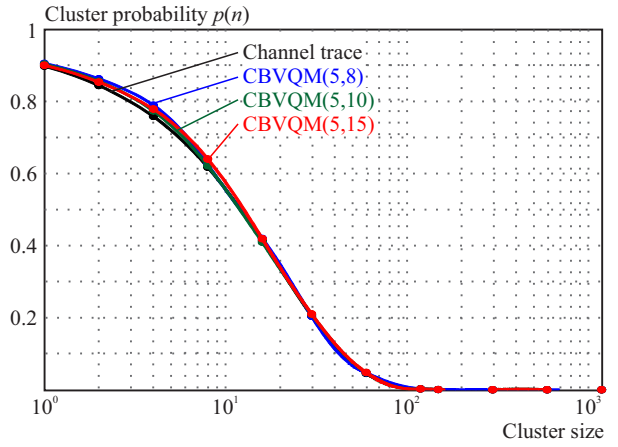

Fig. 13. Cluster error probability curves p(n) for selected CBVQM

Table 1 contains the results of distance analyses on the burst error generated data. As expected after observing the cluster error probability and the error burst distribution, the HVQ-BE model did not pass the χ^2 goodness of fit test and could not even hypothetically be considered sufficient to model the burst error process. However, the other versions both pass the goodness of fit test, hence can be considered a hypothetically sufficient quality models of the observed binary burst error process.

These conclusions are also confirmed by the SE_i characteristic (Fig. 8), which confirms high variance of the observed and modeled process in case of the HVQ-BE version, but extremely low variance through the entire observed interval for the other versions of the HVQ corresponds with previous PDF and CDF observations about high precision. Well representation of the reference stochastic process using the HVQ-BG and HVQ-D versions and absolute failure of fitting the process by the HVQ-BE model can be visually confirmed by the P-P plot in Fig. 9.

3) Error gap analysis

Error gap model data of the observed binary error gap process shows higher difference between the two favored gap multiresolution versions of the HVQ, but surprisingly, even the first model with higher resolution on error bursts produced interesting results. Observations of the Fig. 10 leads to a conclusion, that visual inspection can only hardly determine a superior technique, therefore a closer look at the results of the distance analyses is necessary.

As can be seen from the distance metrics, the well-balanced nature of the modeled error gap process prohibiting visual establishment of superior modeling approach manifested itself in the form of well-balanced distance metrics that could not conclusively establish a winner. Although all the HVQ models could be considered for real error gap modeling after fulfilling the null hypothesis and passing a goodness of fit test, the real winners are again the HVQ-BE (MSE) and HVQ-D (JD, HD and χ^2) models with more distance metrics in favor of the latter.

The well-balanced error gap values produced by all models produce a very small SE overall, and the HVQ-

Table 3. Distance metric evaluation of probability $p(n)$

	JD (D_J)	HD (D_H)	MSE
CBVQM(5,8)	0.00022	0.01584	0.00148
CBVQM(5,10)	0.00008	0.00988	0.00047
CBVQM(5,15)	0.0002	0.01528	0.0011

Table 4. Distance metric evaluation of error burst and gap fit

	Error burst				Error gap			
	JD (D_J)	HD (D_H)	MSE	$C(\chi^2)$	JD (D_J)	HD (D_H)	MSE	$C(\chi^2)$
CBVQM(5,8)	0.00903	0.10331	0.01281	0.01259	0.00826	0.09822	0.00383	0.00024
CBVQM(5,10)	0.02559	0.19682	0.00525	0.09887	0.02467	0.16964	0.01274	0.00043
CBVQM(5,15)	0.00507	0.14562	0.01813	0.01308	0.08792	0.33198	0.02020	0.01251

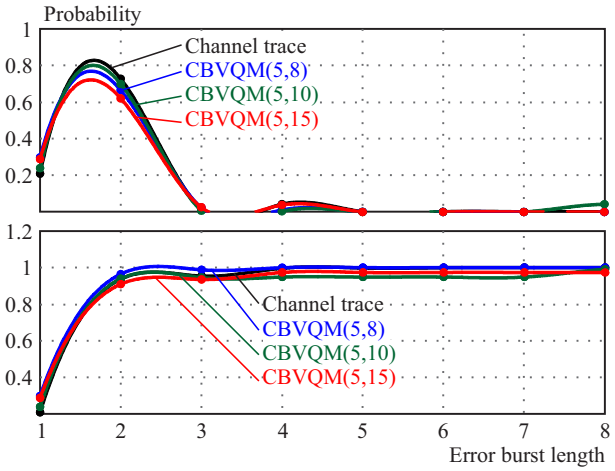

Fig. 14. Error burst histogram (up) and CDF (down) for all models and the real channel data

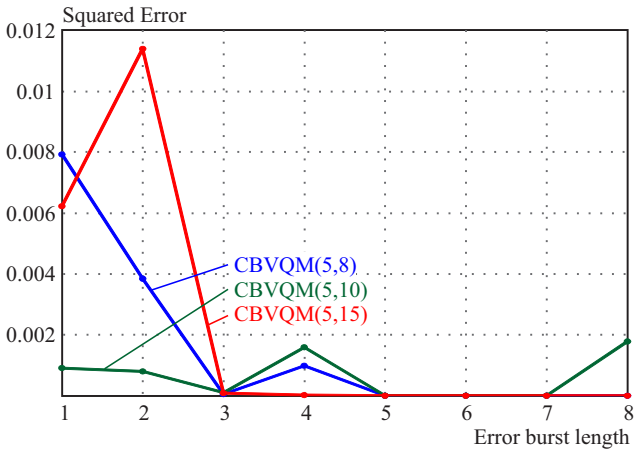

Fig. 15. Error burst se curves for selected CBVQM

BE most probably boasts a better MSE distance solely because of the better fit for gap length 2. The SE continually fades until the gap length 11, after which point it is almost unobservable. Furthermore, the overall variance of the error as depicted on the vertical axis is in reality very small, in the order of 10^{-3}, an extremely high precision for a model.

A well-balanced fir of the optimal line and proposed model in P-P plot (Fig. 12) for the favorited two models makes it difficult to establish a clear winner just by visual inspection. However, all models offer a relatively good fit, with HVQ-BG and HVQ-D demonstrably more precise results.

6.2 CBVQM

1) Cluster error analysis

Cluster error probability (probability of an error-free cluster of a particular size) analysis for a 5 class variant of the CBVQM (Fig. 13) demonstrates a very good fit for every vector length present. The results suggest the presence of such a factor as optimal combination of both the number of classes and the vector length determine model precision.

Quantification of the 5 class CBVQM variant models (Tab. 3) shows that the best model fit to the cluster data from the trace observation is achieved by the variant using vector length 10. However, it has to be added, that all models produce an excellent fit, not just the CBVQM(5,10), but the combination of parameters appears to be the most suitable choice for this particular modeling problem.

2) Error burst analysis

Clearly superior version of the model can once again not be established by simple visual inspection of the observed and generated error burst PDF and CDF (Fig. 14).

All proposed models seem equally capable of producing a sufficiently precise data set, therefore the distance analysis is performed to obtain clear results (Tab. 4).

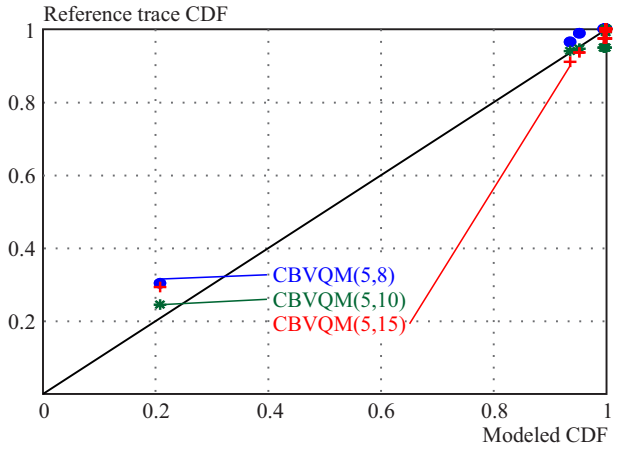

Fig. 16. P–P plot of the error burst models compared to the reference trace

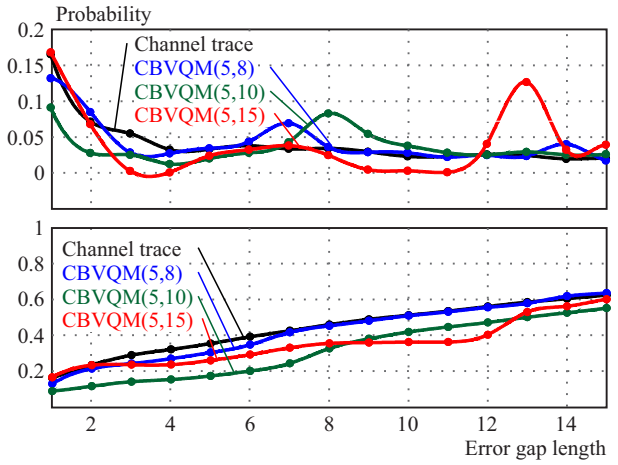

Fig. 17. Error gap histogram (up) and CDF (down) for all models and the real channel data

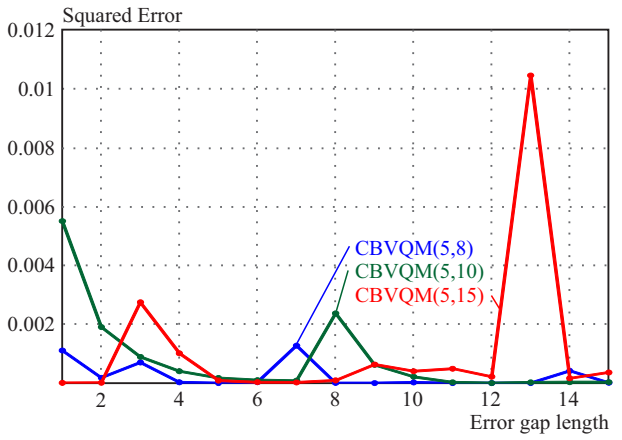

Fig. 18. Error gap SE curves for selected CBVQM

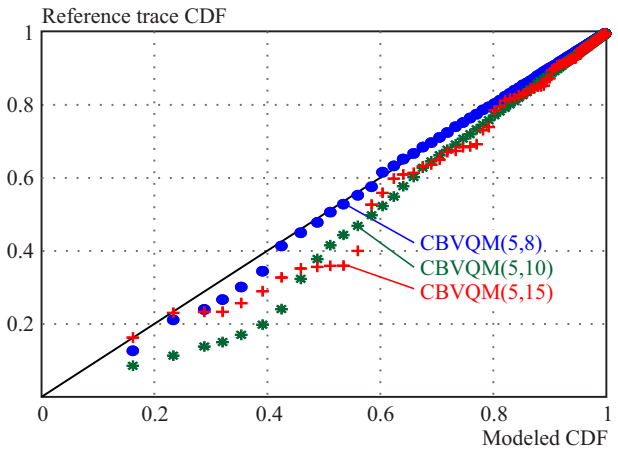

Fig. 19. P-P plot of the error burst models compared to the reference trace

The data clearly shows that the 5 class version of the CBVQM with generating vector length 10 bits does not produce the desired results. It is possible and observation from different runs confirm, that the classification process created a centroid from the binary trace that allows inclusion of vectors capable of degrading the output characteristic to a degree that makes it impossible for the model to pass the goodness of fit test.

This leads to a conclusion that a revision should be taken at some point to improve and optimize the codebook vector selection process to avoid including potentially harmful vectors. One of the possible variants is the unrealized parameterization proposition presented in the model description section.

Squared errors for the CBVQM variants can be seen in Fig. 15. The error is most dominant in the shorter burst lengths and slowly diminishes as it passes to the bigger lengths.

The P-P plots of the analyzed CBVQM variants (Fig. 16) also demonstrate a very good fit of all models to the ideal diagonal line.

3) Error gap analysis

Error gap analysis of the results produced by the CBVQM for 5 classes produces results of gap process distribution that can be seen in Fig. 17. From the first glance it can be seen that each of the modeled sequences fits the desired observed PDF well, with the exception of several local maxima that alter the overall statistics.

These produced maxima further prove that the chosen parameterization method is not yet perfect, because it can include in the codebook such vectors that would not be observed in the real binary trace.

Interestingly, even despite the obvious local minima and maxima, the overall statistic of the data is not affected to a degree that would cause a fail of the goodness of fit test (Tab. 4). Furthermore, it can be seen that the CBVQM achieves very good results, even when compared to the previous models and, more importantly, unlike the empirical models and the reference Elliot's model [25], both the burst error and error gap statistics pass the goodness of fit test, therefore the model as a whole can be confidently used.

The SE_i of the observed process (Fig. 18) behaves as expected after identifying the local maxima already observed in the PDF (Fig. 7). Apart from low gap lengths, the highest squared errors occur at gap lengths approximately corresponding to the vector sizes used in the respective models. This fact reinforces the conclusion that substandard vectors must still be passing the stringent distance criteria introduced in the parameterization process and get to the codebook.

The P–P plot (Fig. 19) demonstrates high precision of the CBVQM(5,8) model further confirming the assumption that the shorter vector lengths yield more precise results for this particular application of the proposed novel classification based VQ model.

7 Conclusion

This paper introduced a novel model concept for arbitrary digital channel error process. Such models are typically used for a variety of tasks, a great example could be replacing the application of binary symmetric channel in [27] to produce an even better specialization for a chip implementation. Mobile network link design is another particularly interesting application currently being explored.

The proposed model concept is based on vector quantization with codebook, constructed using two different techniques. First one represents utilization of standard vectors contained in different orders of Hadamard matrices used for their beneficial purposes, and the second technique introduces a different concept allowing for arbitrary VQ modulating Markov chain structure. Furthermore, the presented novel concept's applicability to modeling a real trace was verified using the goodness of fit tests and compared using the statistical distances.

The randomness is into the VQ-based models introduced by a DTMC, where states are entities representing groups of vectors within the codebook.

Results presented in Tabs. 2 and 4 clearly demonstrate applicability of the models to modeling the wireless digital channel error trace.

HVQ model is, given its various modifications, a superior model of the two propositions. Thanks to the already mentioned beneficial properties of the vectors contained in Hadamard matrices, the channel error trace modeling capability of the HVQ achieves better overall results than the second proposed alternative, the CBVQM. Utilization of a different standard vector set for the codebook construction can lead to even more superior results.

As such, the CBVQM offers an innovative approach to optimal codebook construction, but factors such as the optimal number of classes and vector length remain a key issue. Because this problem cannot be solved analytically, different models were constructed using different vector length (8, 10 and 15) and 5 classes. Further analysis of optimal class number and vector length is out of the scope of this paper. However, as the results in the previous section confirm that shorter vectors should be preferred in real applications to increase model precision. This is caused primarily by inclusion of low quality vectors from the entire alphabet of vectors with the specified length in the initialization phase of the generation process. A more elaborate approach using a different distance punishing rule or vector selection could be applied to improve the precision of the model and quality of its output.

Results with different numbers of classes were also performed, but they do not contradict the findings obtained from the results and observations presented in this paper.

As for the results, two of the presented CBVQM were successful at passing the goodness of fit tests for both their error burst and gap process. Even more, they exhibit extremely good fit of the cluster error probability. They are a viable and efficient replacement or alternative to the unsuitable empirical models and insufficiently precise Elliot's model [25]. The universal nature of codebook construction also predetermines application of this model to any realistic binary modeling problem to which it can naturally and efficiently adapt, removing much of the limitations binding other model concepts.

Conclusively, both model types, the HVQ and CBVQM are more than a viable alternative to the current state-of-the-art models, such as the Elliot's model.

Acknowledgements

The research presented in this paper was financially supported by VEGA 1/0789/15.

References

[1] L. N. Kanal and A. R. K. Sastry, "Models for Channels with Memory and their Applications to Error Control", *Proceedings of the IEEE*, vol. 66, no. 7, pp. 724–744, July 1978.

[2] E. N. Gilbert, "Capacity of a Burst-Noise Channel", *Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, September 1960.

[3] E. O. Elliot, "Estimates of Error Rates for Codes on Burst-Noise Channels", *Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, September 1963.

[4] A. Willig, "A New Class of Packet- and Bit-Level Models for Wireless Channels", *The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 5, pp. 2434–2440, 2002.

[5] T. Tao, J. Lu, and J. Chuang, "Hierarchical Markov Model for Burst Error Analysis Wireless Communications", *IEEE VTS 53rd Vehicular Technology Conference*, vol. 4, pp. 2843–2847, 2001.

[6] J. Polec, V. Hirner, M. Martinovič, K. and K. Kotuliaková, "A Generator from Cascade Markov Model for Packet Loss and Subsequent Bit Error Description", *World Academy of Science, Engineering and Technology*, vol. 7, no. 4, 2013.

[7] W. Turin and M. Sondhi, "Modeling Error Sources Digital Channels", *IEEE Journal on Selected Areas Communications*, vol. 11, no. 3, pp. 340–347, 1993.

[8] O. S. Salih, C.-X. Wang and D. I. Laurenson, "Double Embedded Processes Based Hidden Markov Models for Binary Digital Wireless Channels", *IEEE International Symposium on Wireless Communication Systems*, pp. 219–223, 2008.

[9] C. Jiao, L. Schwiebert and B. Xu, "On Modeling the Packet Error Statistics Bursty Channels", *27th Annual IEEE Conference on Local Computer Networks*, pp. 534–541, 2002.

[10] M. U. Ilyas and H. Radha, "A Channel Model for the Bit Error Rate Process 802.15. 4 LR-WPAN Wireless Channels," *Proceedings of IEEE International Conference on Communications*, pp. 257–261, 2008.

[11] A. Nogueira, P. Salvador and R. Valadas, "Fitting Algorithms for MMPP ATM Traffic Models", *Proceedings of the Broadband access conference*, 1999.

[12] A. Nogueira and R. Valadas, "Analysing the Versatility of the 2-MMPP Traffic Model", *Proceedings of the Second International Symposium on Communication Systems Networks and Digital Signal Processing*, 2000.

[13] S. H. Kang, Y. Kim, D. K. Sung and B. D. Choi, "An Application of Markovian Arrival Process (MAP) to Modeling Superposed ATM Cell Streams", *IEEE Transactions on Communications*, vol. 50, no. 4, pp. 633–642, 2002.

[14] H. Okamura, T. Dohi and K. S. Trivedi, "Markovian Arrival Process Parameter Estimation with Group Data", *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 4, pp. 1326-1339, 2009.

[15] P. Pruthi and A. Erramilli, "Heavy-Tailed on/off Source Behavior and Self-Similar Traffic", *IEEE International Conference on Communications*, vol. 1, pp. 445–450, 1995.

[16] E. Costamagna, L. Favalli, P. Gamba and P. Savazzi, "Block-Error Probabilities for Mobile Radio Channels Derived from Chaos Equations", *IEEE Communications Letters*, vol. 3, no. 3, pp. 66–68, 1999.

[17] E. Costamagna, A. Fanni, L. Favalli and P. Gamba, "Experiments Modeling the Parameters of Chaos Equation Models for Mobile Radio Channels", *IEEE-APS Conference on Antennas and Propagation for Wireless Communications*, pp. 103–106, 2000.

[18] A. Kopke, A. Willig and H. Karl, "Chaotic Maps as Parsimonious Bit Error Models of Wireless Channels", vol. 1, pp. 513–523, 2003.

[19] C. X. Wang, W. Xu and M. Pätzold, "Error Models for Evaluating Error Control Strategies EGPRS System", IEEE 60th Vehicular Technology Conference, vol. 6, pp. 4238–4244, 2004.

[20] C. X. Wang and W. Xu, "A New Class of Generative Models for Burst-Error Characterization Digital Wireless Channels", *IEEE Transactions on Communications*, vol. 55, no. 3, pp. 453–462, 2007.

[21] C. X. Wang and M. Pätzold, "A new Deterministic Process based Generative Model for Characterizing Bursty Error Sequences", Proceedings IEEE, 2004.

[22] S. A. Khayam, H. Radha, S. Aviyente and J. R. Deller Jr., "Markov and Multifractal Wavelet Models for Wireless MAC-to-MAC Channels", *Performance Evaluation*, vol. 64, no. 4, pp. 298–314, 2007.

[23] R. Ranjan, D. Bepari and D. Mitra, "Genetic Algorithm based Finite State Markov Channel Modeling", *International Journal of Wireless Communications and Mobile Computing*, vol. 1, no.4, pp. 96–102, 2013.

[24] T. Csóka and J. Polec, "Analysis of Additive Noise Characteristics Indoor Wireless Sensor Networks", *EUROCON 2015 – International Conference on Computer as a Tool (EUROCON). IEEE*, 2015.

[25] T. Csóka, J. Polec, I. Ilčíková and J. Doboš, "Binary Error Models for Wireless Sensor Networks", *IWSSIP 2016: The 23rd International Conference on Systems*, Signals and Image Processing.Bratislava, Slovakia.ISBN 978-1-4673-9554-0.23-25 May 2016.

[26] M. Bella "Miery dobrej zhody založené na štatistických vzdialenostiach", *Bratislava*, 2013.

[27] T. Páleník, P. Farkaš and M. Rakús, "Analysis of Minimal LDPC Decoder System on a Chip Implementation", *Radioengineering*, vol. 24, no. 3, pp. 783–790, 2015.

Received 3 August 2016

Jaroslav Polec received the Ing and PhD degrees in telecommunication engineering from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in 1987 and 1994, respectively. Since 1997 he has been associate professor and since 2007 professor at the Institute of Multimedia Information and Communication Technologies of the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology and since 1998 at the Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics of Comenius University. His research interests include automatic repeat request, channel modeling, image coding, interpolation, and filtering.

Tibor Csóka received his PhD degree in telecommunication engineering from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, in 2016 and is currently a software developer for NOKIA. His interests include channel modeling, signal processing for communications, image processing, and radar processing.

Kvetoslava Kotuliaková was born in 1968 in Bohumín, Czech Republic. She received the MS degree in telecommunication engineering from the Faculty of Electrical and Information Technology, Slovak University of Technology in 1992 and PhD degree in 2005 at the same university. From 2013 she is an associate professor at Institute of Multimedia Information and Communication Technologies of the Faculty of Electrical and Information Technology, Slovak University of Technology. Her research interests include error control, channel modelling and traffic.

Filip Csóka, is currently a PhD student at the Institute of Multimedia Information and Communication Technologies, FEI STU in Bratislava. He focuses on digital signal processing and digital image processing in his pedagogic activities. In his scientific research he focuses on image recognition, mainly sign recognition. He has experience working with signaling in VoIP systems and had designed various security mechanisms for VoIP systems.