



**TD(06)007**

**5th COST 290 Management Committee and Technical Meeting**

**Delft, February 9-10, 2006**

**Impact of SLA to VoIP QoS**

Martin Klimo, Juraj Smiesko  
University of Zilina  
Univerzitna 1, 010 26 Zilina, Slovakia  
{martin.klimo, juraj.smiesko}@fri.utc.sk

**Abstract**

It is well known, that a bitrate of speech at a telephone quality is 64 kbit/s. Capacity saving can be reached by silence suppression using a voice activity detection (VAD). In this case, speech stream cannot be described in SLA just by this peak-rate, but some burstiness descriptor has to be used also. The paper explains, how to plan values of the token bucket control mechanism to reach maximum of voice quality if VAD is applied. For proper modeling of the control mechanism, also the model of speech with VAD has to be introduced.

**Keywords**

VoIP dimensioning, QoS, Token Bucket

**Working Group 3**

# Introduction

While in telephone networks intelligence is located in network nodes, intelligence in computer networks is much more dedicated to end devices. Even in the digital telephone network, the role of the end device is just to make a call and transform an acoustical pressure to an electrical signal. Processor based end devices allow to process acoustical signal before it's sending to the network. One of the main tasks of speech processing is removing of signal redundancy that leads to the network capacity saving. Well known kind of such as processing is a speech compression that allows transmission speed reduction in channel switched networks. Packet switched networks allow additional capacity savings using voice activity detection (VAD). VAD fully or partially suppress silence periods that reduces amount of transmitted data. On the other side, VAD changes a type of the voice stream. While digitalised speech is a constant bitrate stream with the same mean rate as a peak rate i.e. 64 kbit/s, if VAD is applied, speech stream consists of bursts (talk-spurts) and silent periods (no background noise. transmission is assumed in the paper). This has a strong impact to the Service Level Agreement (SLA), as well as to the access network dimensioning. The voice stream cannot be described by one parameter anymore, and the stream descriptors depend not only on voice properties but also on a policy algorithm. Token bucket algorithm is assumed in the paper. The biggest advantage of this policing strategy is no additional delay, and SLA overrunning leads only to packet loss. This fits in with the fact, that IP telephony is more delay sensitive than loss sensitive service. Nevertheless packet loss degrades quality of the service, and the aim of the paper is to show, how to set an average token rate and token bucket depth to save link capacity on an acceptable speech quality.

The paper is organised as follows: section 2 describes the speech stream model without and with VAD. Section 3 presents a model of token bucket supplied by On-Off speech stream. Section 4 introduces speech quality evaluation used in the paper, and shows relation between capacity savings and speech quality.

## 1 Markov model of VoIP stream

TASI system introduced an idea that silent periods in telephony dialog can be used for link capacity savings. Using the same principle in IP telephony brings the same savings, but not only on undersea cables but on the whole network. Unlike of constant bit rate speech stream in digital telephony network, speech with applied VAD has different structure which has to be understudied, and proper models have to be developed.

### 1.1 Speech stream

First of all it is necessary to know talk-spurt and silent period distribution in a natural dialog. Because this process is a basic workload of telephony network, this problem is well studied and the corresponding model is even standardised [1]. We use this model as a starting point.

The On-Off Source Model of human speech is composed of two periods, where On period belongs to the active periods in human speech (talk-spurt) and the Off period belongs to the silence (gap). In any time instant the actual period may be finished and the state will be switched to the different state. The talk-spurt duration is modelled by the stochastic variable  $T_1$  and the pause duration is modelled by the variable  $T_2$ . The probability densities of  $T_1$  and  $T_2$  are modelled by two weighted geometric distribution functions. Every increment of these

variables  $T_1$  and  $T_2$  is equal to 5 ms and then the average talk-spurt duration is  $ET_1 = 227$  ms and the average pause duration is  $ET_2 = 596$  ms (see [1]).

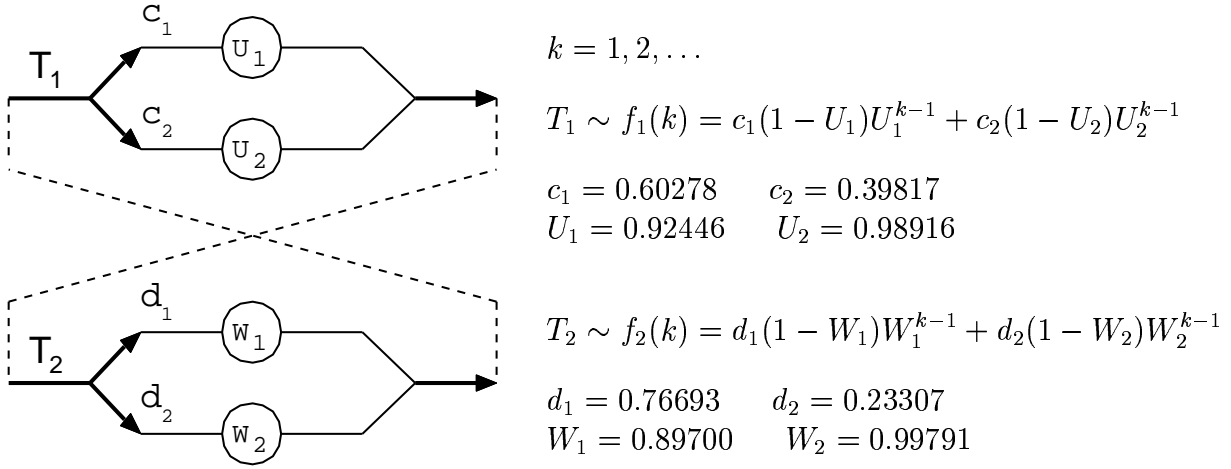


Figure 1: Hypergeometric model

This model is called Hypergeometrical and one example of the process behaviour is on Fig.2:

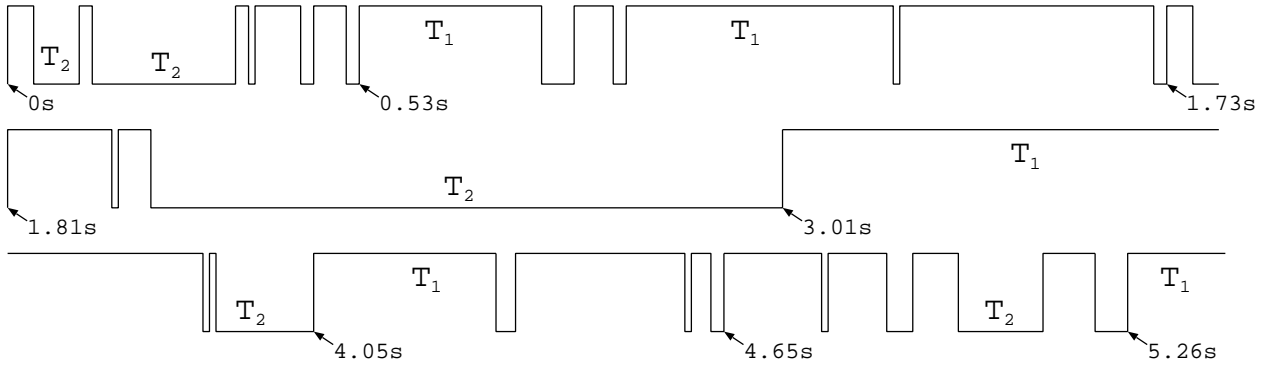


Figure 2: Active/silence periods in 5.4s speech segment

We are using continuous time models within this paper. The best continuous approximation of discrete hypergeometrical model is the hyperexponential model. The discrete geometrical distribution are approximated by continuous exponential ones and the rest of this model has completely the same nature as the geometrical one. In the new model we keep the same averages.

$$\text{ON: } T_1 \sim f_1(t) = c_1\alpha_1 e^{-\alpha_1 t} + c_2\alpha_2 e^{-\alpha_2 t}$$

$$c_1 = 0.60278 \quad \alpha_1 = 0.01511 \text{ ms}^{-1} \quad c_2 = 0.39817 \quad \alpha_2 = 0.00217 \text{ ms}^{-1}$$

$$\text{OFF: } T_2 \sim f_2(t) = d_1\beta_1 e^{-\beta_1 t} + d_2\beta_2 e^{-\beta_2 t}$$

$$d_1 = 0.76693 \quad \beta_1 = 0.02060 \text{ ms}^{-1} \quad d_2 = 0.23307 \quad \beta_2 = 0.00042 \text{ ms}^{-1}$$

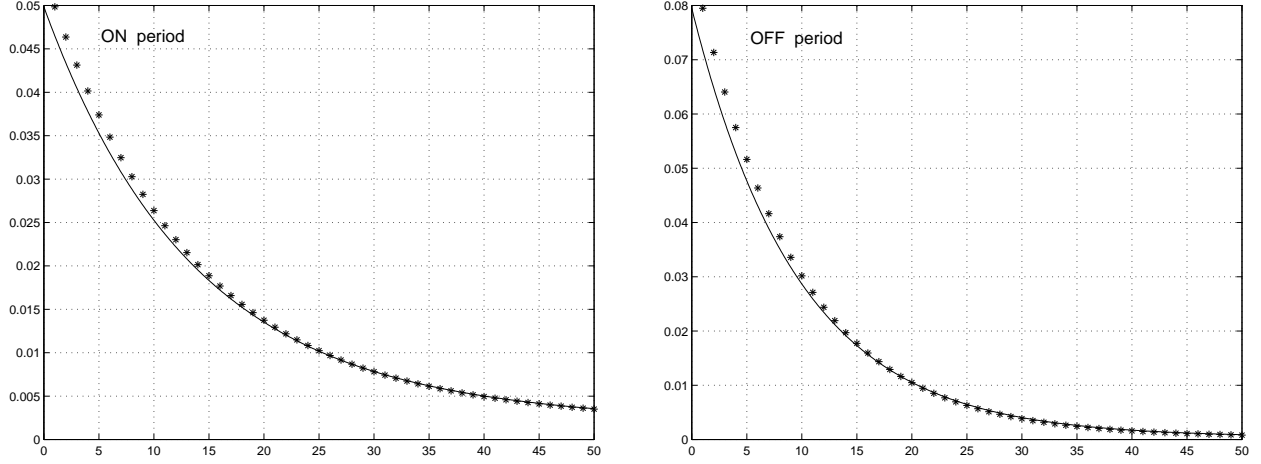
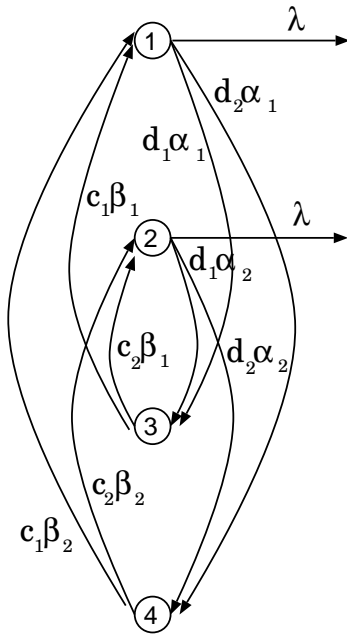


Figure 3: Approximation by hyperexponential distribution for On and Off periods

The packet voice traffic will be generated by this four-states Hyperexponential source (see Fig.4), where states 1 and 2 represent On period and states 3 and 4 represent Off period. When the source is in On state it starts generating flow of packets with intensity  $\lambda$ . This source is just an Markovian approximation of real codec, which generates packets during On period regularly. In following examples we assume the packet rate (using G.729A)  $\lambda = 50$  p/s. We denote Hyperexponential source as HM<sub>2</sub>-HM<sub>2</sub> (HM<sub>2</sub> for two-weight hyperexp. distribution). Fig.4 shows a transition diagram of the related Markov chain source (loops are omitted):



The rate matrix of this Markov chain is  $\mathbf{Q}_0$ :

$$\mathbf{Q}_0 = \begin{pmatrix} -\alpha_1 & 0 & d_1\alpha_1 & d_2\alpha_1 \\ 0 & -\alpha_2 & d_1\alpha_2 & d_2\alpha_2 \\ c_1\beta_1 & c_2\beta_1 & -\beta_1 & 0 \\ c_1\beta_2 & c_2\beta_2 & 0 & -\beta_2 \end{pmatrix}$$

Stationary probability of the chain  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$  is given by the balance equations  $\boldsymbol{\pi} \cdot \mathbf{Q}_0 = \mathbf{0}$  and the normalisation condition  $\sum_{\forall i} \pi_i = 1$ :

$$\begin{aligned} \pi_2 &= \frac{c_2\alpha_1}{c_1\alpha_2} \pi_1 & \pi_3 &= \frac{d_1\alpha_1}{c_1\beta_1} \pi_1 & \pi_4 &= \frac{d_2\alpha_1}{c_1\beta_2} \pi_1 \\ \pi_1 &= \frac{c_1\alpha_2\beta_1\beta_2}{(c_1\alpha_2 + c_2\alpha_1)\beta_1\beta_2 + (d_1\beta_2 + d_2\beta_1)\alpha_1\alpha_2} \end{aligned}$$

Figure 4: Markov chain for HM<sub>2</sub>-HM<sub>2</sub> source

$P_{ON} = \pi_1 + \pi_2$  and  $P_{OFF} = \pi_3 + \pi_4$  are probabilities that the source is in state On or Off. The probabilities for parameters of VoIP given above are

$$\boldsymbol{\pi} = ( 0.04902, 0.22546, 0.04579, 0.67973 ), \quad P_{ON} = 0.27448, \quad P_{OFF} = 0.72552$$

Mean  $\eta$  of packets generated by source is  $\eta = \lambda \cdot P_{ON} = 13.7$  p/s

## 1.2 Speech + VAD stream

Even when highest capacity savings can be reached if any silence in the speech is not sampled and transmitted, total silence suppression is not the best strategy. Clipping, adaptive jitter buffering instabilities and comfort noise generating during short gaps, which are identified by VAD, may bring quality degradation that is not compensated by negligible link capacity reduction. This is why the better VAD strategy may be leaving short gaps for sampling and subsequent transmission. A limit for minimal gap duration depends on method for silence detection and setting of its parameters. We assume in the paper that Voice Activity Detector allows setting a fixed but configurable limit for a minimum gap that will be suppressed. There are two basic techniques for silence suppression [2]: hangover and fill-in. A hangover adds at the end of spurt a fixed interval that prolongs this spurt and signal within this interval is sampled. If no speech is detected at the end of the interval, the gap is detected and samples are suppressed. Implementation of this algorithm is simple, but there is no guarantee that short silent periods will not occur. The silent sates in the correspondent model are same as for pure speech (see Fig.4) only routing probabilities  $c_1$  and  $c_2$  will differ. A fill-in mechanism bridges all gaps belonging to the given limit and remains gaps that are longer then the fill-in time. To do that, a look-ahead delay is necessary, which degrades speech quality again. To study the impact of an short gap suppression while long gaps are preserved, fill-in mechanism was chosen for the paper.

Previous studies on On-Off period distributions (see [3]) are based on On-Off periods measurement of a voice stream at the output of voice activity detector and fitting parameters of an appropriate model. Our approach is based on the speech model standardised by ITU-T as it was described in the previous section and the impact of VAD is derived from the silence suppression by the fill-in mechanism. We assume that the limit is configurable (unfortunately this is not the case of G729.B VAD). Paper shows three examples, when the fill-in time is set to 20 ms, 200 ms, and 1 s. The speech at the VAD output is called in the paper as a "modified speech". It is modelled by On-Off source again, only the spurt and gap period distributions should be modified. Let the talk-spurt duration is modelled by variate  $X$ , and the pause duration is modelled by variate  $Y$ .

Fig.5 gives an example if, in speech given by Fig.4, silence shorter than 20 ms is not suppressed.

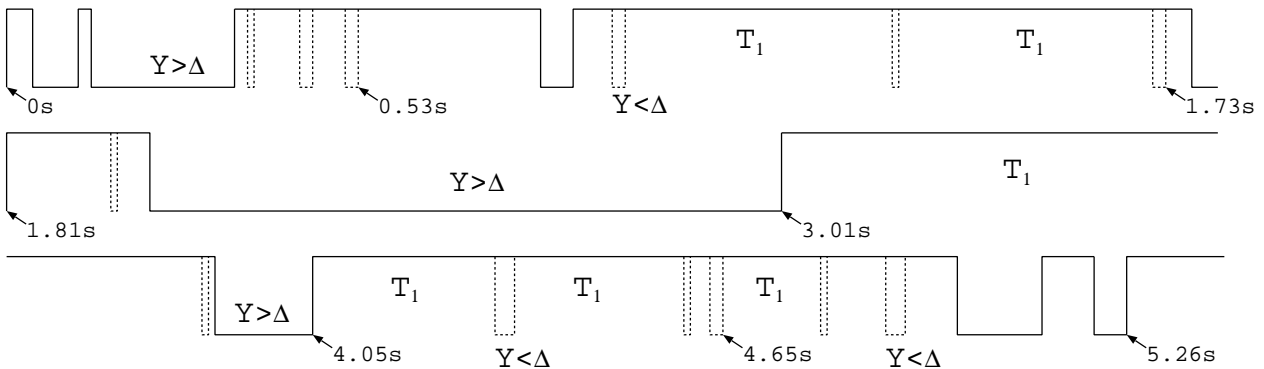


Figure 5: Active/silence periods - silence shorter than 20 ms is not suppressed.

Our aim is to find probability distributions of new stochastic variables variates  $X$  and  $Y$ .

Let  $p$  is probability of OFF period is shorter than  $\Delta$ :

$$p = P(T_2 < \Delta) = d_1 (1 - e^{-\beta_1 \Delta}) + d_2 (1 - e^{-\beta_2 \Delta}) = 1 - [d_1 e^{-\beta_1 \Delta} + d_2 e^{-\beta_2 \Delta}]$$

Let us obtain probability distribution of stochastic variable  $X$  for On period. First, we assume an existence of the contiguous  $(k - 1)$  "short" OFF periods, where  $T_2 < \Delta$ . The probability of this random event is  $(1 - p)p^{k-1}$ . Then, the stochastic variable  $X$  is a sum of  $k$  stochastic variables  $T_1$  and  $(k - 1)$  variables  $T_2$  (\* mean convolution sum, see Fig.6):

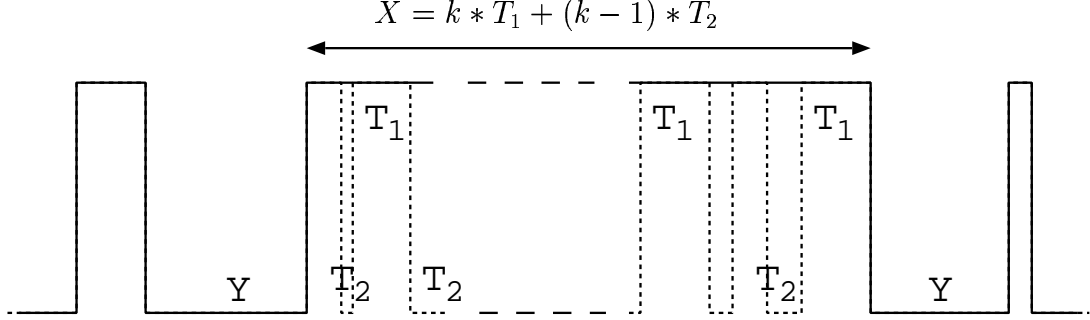


Figure 6: Convolution relation for variate  $X$

Probability density function of variate  $X$  will be given by following convolution relation, when an independent assumption is used:

$$f_X(t) = \sum_{k=1}^{\infty} (1 - p)p^{k-1} [f_1^k(t) * f_2^{k-1}(t)] \quad \text{and} \quad EX = \frac{1}{1 - p} [ET_1 + pET_2]$$

For finding  $f_X(t)$  form we can use the Laplace transformation (LT). Let the LTs of density functions  $f_1(t)$  and  $f_2(t)$  are  $L_1(x) = L\{f_1(t)\}$  and  $L_2(x) = L\{f_2(t)\}$ . The LT of density function  $f_X(t)$  is (we suppose, that  $L_1(x)$  and  $L_2(x)$  satisfies convergence conditions):

$$L_X(x) = \sum_{k=1}^{\infty} (1 - p)p^{k-1} L_1^k(x) L_2^{k-1}(x) = (1 - p)L_1(x) \sum_{k=0}^{\infty} [pL_1(x)L_2(x)]^k = \frac{(1 - p)L_1(x)}{1 - pL_1(x)L_2(x)}$$

Where Laplace transformations of density functions  $f_1(t)$  and  $f_2(t)$  are:

$$L_1(x) = L\{f_1(t)\} = c_1 \frac{\alpha_1}{\alpha_1 + x} + c_2 \frac{\alpha_2}{\alpha_2 + x}, \quad L_2(x) = L\{f_2(t)\} = d_1 \frac{\beta_1}{\beta_1 + x} + d_2 \frac{\beta_2}{\beta_2 + x}$$

Laplace transformation of density function  $f_Y(t)$  is:

$$L_X(x) = \frac{(1 - p) \left[ c_1 \frac{\alpha_1}{\alpha_1 + x} + c_2 \frac{\alpha_2}{\alpha_2 + x} \right]}{1 - p \left[ c_1 \frac{\alpha_1}{\alpha_1 + x} + c_2 \frac{\alpha_2}{\alpha_2 + x} \right] \left[ d_1 \frac{\beta_1}{\beta_1 + x} + d_2 \frac{\beta_2}{\beta_2 + x} \right]} = \frac{(1 - p)P(x)}{Q_4(x) - p \cdot Q_2(x)}$$

$$P(x) = (\beta_1 + x)(\beta_2 + x)(c_1 \alpha_1 (\alpha_2 + x) + c_2 \alpha_2 (\alpha_1 + x))$$

$$Q_4(x) = (\alpha_1 + x)(\alpha_2 + x)(\beta_1 + x)(\beta_2 + x)$$

$$Q_2(x) = [c_1 \alpha_1 (\alpha_2 + x) + c_2 \alpha_2 (\alpha_1 + x)] [(d_1 \beta_1 (\beta_2 + x) + d_2 \beta_2 (\beta_1 + x))]$$

Let  $k_1, k_2, k_3$  and  $k_4$  are polynomial roots in denominator  $Q(x) = Q_4(x) - p \cdot Q_2(x)$ . Then Laplace transformation  $L_X(x)$  is given by

$$L_X(x) = \frac{(1-p)P(x)}{(x-k_1)(x-k_2)(x-k_3)(x-k_4)} = (1-p) \cdot \frac{P(x)}{Q(x)}$$

Using residual method we can reduce function  $L_X(x)$  to partial fractions:

$$L_X(x) = (1-p) \cdot \frac{P(x)}{Q(x)} = (1-p) \cdot \sum_{i=1}^4 \frac{P(k_i)}{Q'(k_i)} \frac{1}{x-k_i}$$

$$Q'(k_i) = (k_i - k_3)(k_i - k_4) [(k_i - k_2) + (k_i - k_1)] + (k_i - k_1)(k_i - k_2) [(k_i - k_4) + (k_i - k_3)]$$

Original of Laplace transformation is  $f_X(t) = (1-p) \cdot \sum_{i=1}^4 \frac{P(k_i)}{Q'(k_i)} \cdot e^{k_i t}$ . We assign  $r_i = -k_i$

and  $B_i = \frac{(1-p)}{r_i} \cdot \frac{P(k_i)}{Q'(k_i)}$ . We can see, that  $B_i$  are probabilities and  $\sum_{i=1}^4 B_i = 1$ .

We found out, that the stochastic variable  $X$  is modelled by four-weighted hyperexponential distribution function  $HM_4$

$$f_X(t) = \sum_{i=1}^4 B_i \cdot r_i e^{-r_i t} \quad t \in \langle 0, \infty \rangle \quad \text{and} \quad EX = \sum_{i=1}^4 \frac{B_i}{r_i}$$

Probability distribution of variable  $Y$  will be given:

$$f_Y(t) = f_2(t/T_2 > \Delta) = \frac{f_2(t)}{P(T_2 > \Delta)} = \frac{f_2(t)}{1-p} = \frac{d_1 \beta_1 e^{-\beta_1 t} + d_2 \beta_2 e^{-\beta_2 t}}{d_1 e^{-\beta_1 \Delta} + d_2 e^{-\beta_2 \Delta}} \quad t \in \langle \Delta, \infty \rangle$$

The average of variable  $Y$  is  $EY = \frac{d_1 e^{-\beta_1 \Delta} \left[ \Delta + \frac{1}{\beta_1} \right] + d_2 e^{-\beta_2 \Delta} \left[ \Delta + \frac{1}{\beta_2} \right]}{d_1 e^{-\beta_1 \Delta} + d_2 e^{-\beta_2 \Delta}}$

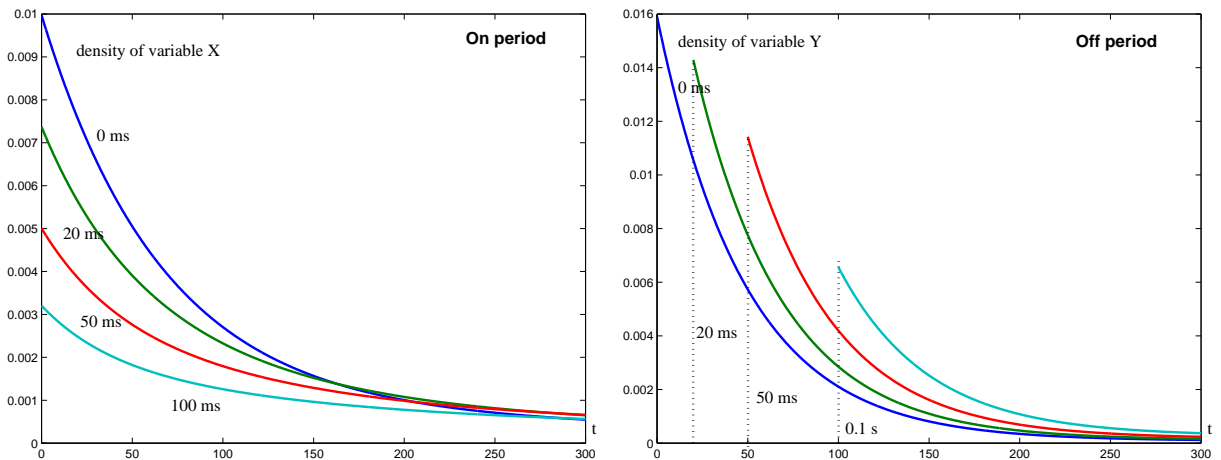


Figure 7: Densities  $f_X(t)$  and  $f_Y(t)$  with  $\Delta = 0, 20ms, 50ms, 100ms$

To underline difference between exponential distributions, which is often used as a basic approximation of the spurt and gaps periods, and approximations used in this paper, Fig.8 shows the complementary gap Cumulative Distribution Function (CDF). The complementary gap CDF are given:

$$F'_X(t) = \sum_{i=1}^4 B_i \cdot e^{-r_i t} \quad t \in \langle 0, \infty \rangle, \quad F'_Y(t) = \frac{d_1 e^{-\beta_1 t} + d_2 e^{-\beta_2 t}}{d_1 e^{-\beta_1 \Delta} + d_2 e^{-\beta_2 \Delta}} \quad t \in \langle \Delta, \infty \rangle$$

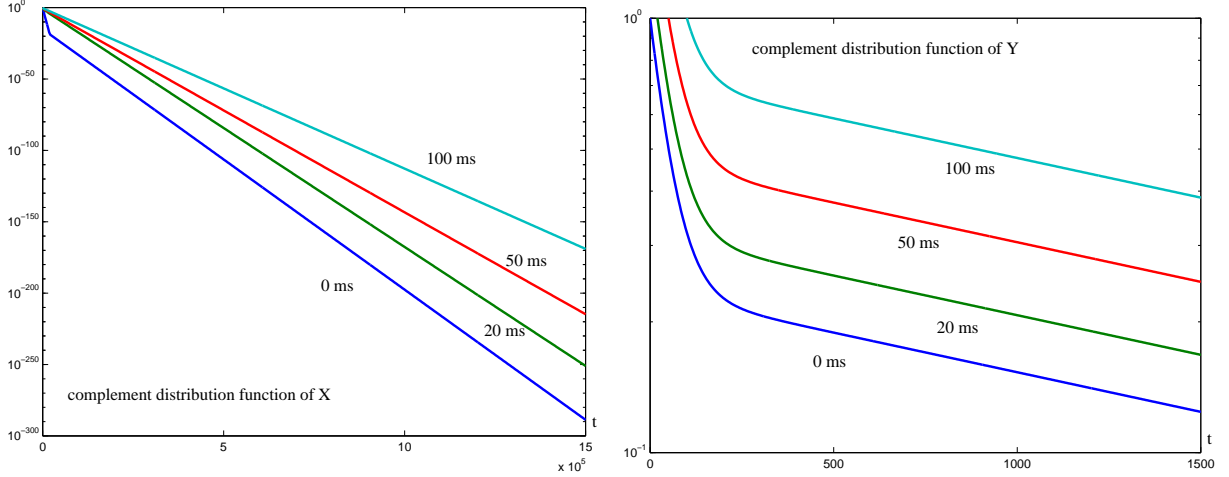


Figure 8: Complement distribution function of  $X$  and  $Y$  with  $\Delta = 0, 20ms, 50ms, 100ms$

The  $y$  axis is plotted in logscale, therefore the complementary CDF of the exponential distributed gap is given by the straight line.

To use theory of Markovov chains we shall approximate  $Y$  by some phases probability distribution. For variation coefficient holds,  $V_Y > 1$ . For this reason we have to use hypererlang distribution for approximation. The most simple is two-weight distribution  $HE_{n_1, n_2}$ :

$$g(t) = p_1 \frac{\mu_1^{n_1} t^{n_1-1} e^{-\mu_1 t}}{(n_1 - 1)!} + p_2 \frac{\mu_2^{n_2} t^{n_2-1} e^{-\mu_2 t}}{(n_2 - 1)!} \quad \text{for } t \in \langle 0, \infty \rangle$$

For estimation of unknown parameters we use the first and second moments of variable  $Y$ :

$$p_1 \frac{n_1}{\mu_1} + p_2 \frac{n_2}{\mu_2} = EY \quad \text{and} \quad p_1 \frac{n_1(n_1 + 1)}{\mu_1} + p_2 \frac{n_2(n_2 + 1)}{\mu_2} = E(Y^2)$$

We have obtained following recurent formulas for parameters:

$$\mu_1 = \frac{p_1 n_1 \mu_2}{\mu_2 EY - p_2 n_2}, \quad \mu_2 = \frac{p_2 (n_1 + 1) n_2 EY + \sqrt{D}}{(n_1 + 1) EY^2 - p_1 n_1 E(Y^2)}, \quad p_2 = 1 - p_1$$

$$D = (n_1 + 1)^2 (p_2 n_2 EY)^2 - p_2 n_2 [(n_1 + 1) EY^2 - p_1 n_1 E(Y^2)] [p_1 n_1 (n_2 + 1) + p_2 n_2 (n_1 + 1)]$$

For given phases  $n_1$  and  $n_2$  we have estimated probability  $p_1$  using minimzing of integral norm  $\| \cdot \|_{\Delta}$  between both density functions:

$$\|f_Y - g\|_{\Delta} = \left[ \int_0^{\Delta} g^2(t) dt + \int_{\Delta}^{\infty} [f_Y(t) - g(t)]^2 dt \right]^{\frac{1}{2}}$$

We show examples of the gap distributions for three values of fill-in time. First value is given by speech frame packed into one packet (we use 20 ms in the paper). Second value is 200 ms, which is used in many Voice Activity Detectors as a default value. And finally fill-in time of 1s gives an upper limit. Due to average silence of 596 ms in a dialog, this limit is not far from a case when no VAD is applied. The following number of phases lead to the most accurate approximations:

$$\begin{aligned} \Delta = 20 \text{ ms} : & \quad n_1 = 1, \quad n_2 = 3, \quad \| \cdot \|_{20} = 0.02488 \\ \Delta = 200 \text{ ms} : & \quad n_1 = 1, \quad n_2 = 2, \quad \| \cdot \|_{200} = 0.00767 \\ \Delta = 1000 \text{ ms} : & \quad n_1 = 3, \quad n_2 = 7, \quad \| \cdot \|_{1000} = 0.00397 \end{aligned}$$

To keep a token bucket model as simple as possible, we have chosen  $HE_{1,3}$ , which is optimal from used measure point of view for fill-in time  $\Delta = 20\text{ms}$ . But also for other filling-times acceptable accuracy was reached:  $\| \cdot \|_{200} = 0.00793$ ,  $\| \cdot \|_{1000} = 0.00561$ .

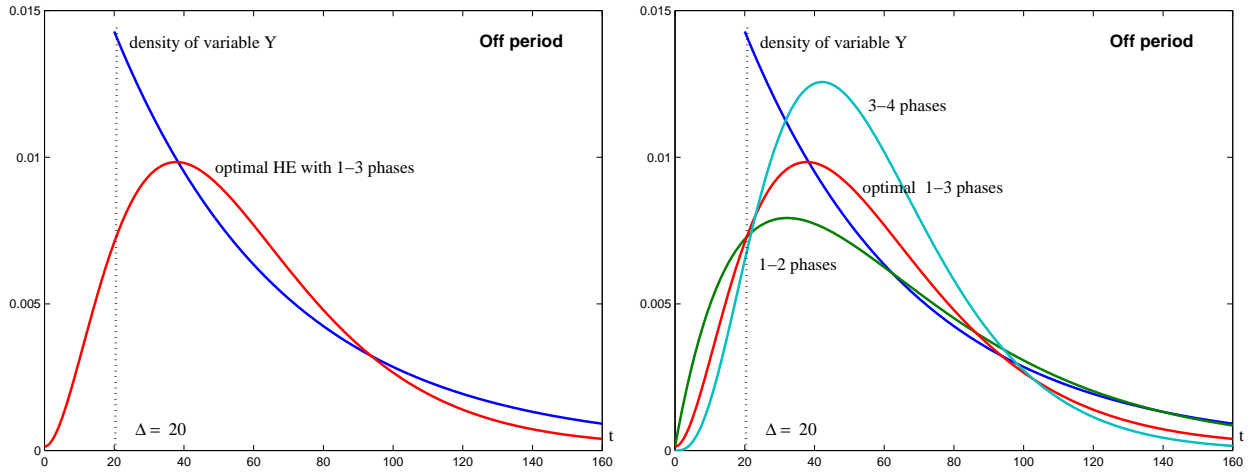


Figure 9: Approximation  $Y$  with  $\Delta = 20 \text{ ms}$  by optimal  $HE_{1,3}$  and  $HE_{1,2}$ ,  $HE_{3,4}$

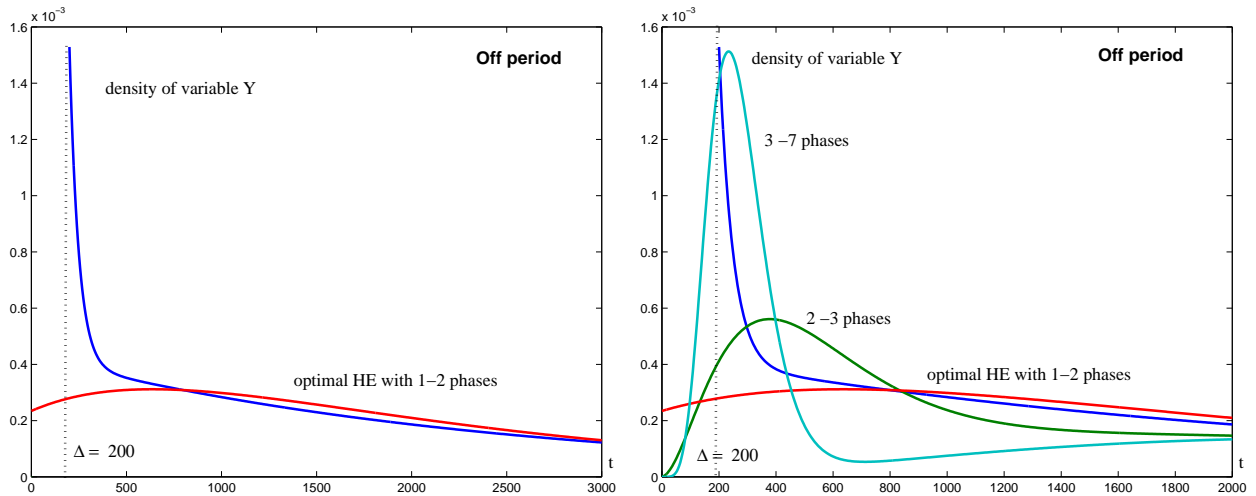


Figure 10: Approximation  $Y$  with  $\Delta = 200 \text{ ms}$  by optimal  $HE_{1,2}$  and  $HE_{2,3}$ ,  $HE_{3,7}$

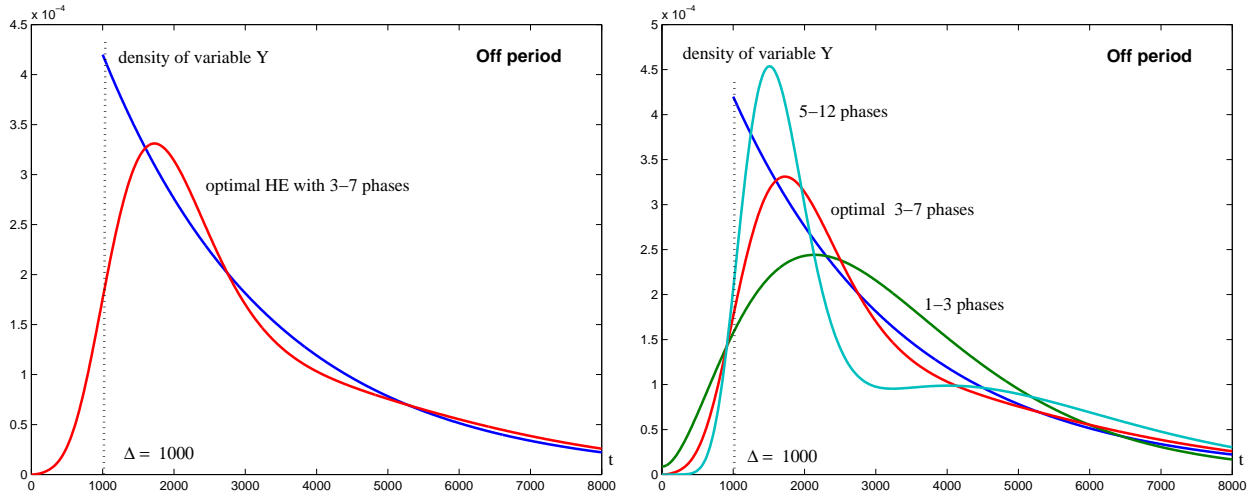
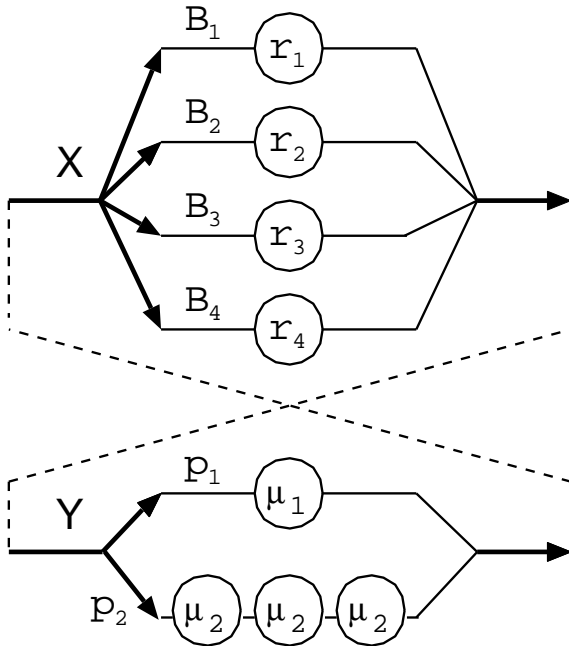


Figure 11: Approximation  $Y$  with  $\Delta = 1000$  ms by optimal  $HM_{3,7}$  and  $HM_{1,3}$ ,  $HM_{5,12}$

To conclude, we have approximated variable  $Y$  (gap) by 2-weight 4-phases hypererlang distribution  $g(t)$  and we have obtained the following Hyperexponential - Hypererlang source  $HM_4$ - $HE_{1,3}$  of human speech with VAD:



**ON:**

$$f_X(t) = \sum_{i=1}^4 B_i \cdot r_i e^{-r_i t} \quad \text{for } t \in (0, \infty)$$

$$EX = B_1 \frac{1}{r_1} + B_2 \frac{1}{r_2} + B_3 \frac{1}{r_3} + B_4 \frac{1}{r_4}$$

**OFF:**

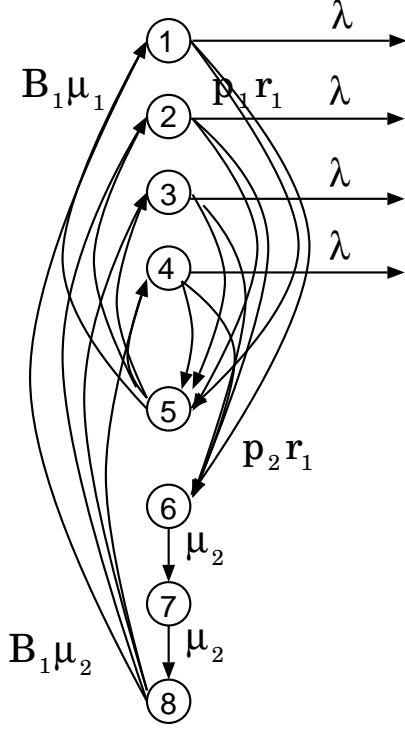
$$g(t) = p_1 \cdot \mu_1 e^{-\mu_1 t} + p_2 \cdot \frac{\mu_2^3 t^2 e^{-\mu_2 t}}{2}$$

$$t \in (0, \infty)$$

$$EY = p_1 \frac{1}{\mu_1} + p_2 \frac{3}{\mu_2}$$

Figure 12: Phase model of speech with partial VAD

Eight-states Markov chain describes behavior of the modified Hyperexponential source described above, where states 1,2,3 and 4 represents On period and states 5,6,7 and 8 represents Off period. Fig.13 shows a transmission diagram of Markov chain:



The rate matrix of this Markov chain is  $\mathbf{Q}_0$ :

$$\begin{pmatrix} -r_1 & 0 & 0 & 0 & p_1 r_1 & p_2 r_1 & 0 & 0 \\ 0 & -r_2 & 0 & 0 & p_1 r_2 & p_2 r_2 & 0 & 0 \\ 0 & 0 & -r_3 & 0 & p_1 r_3 & p_2 r_3 & 0 & 0 \\ 0 & 0 & 0 & -r_4 & p_1 r_4 & p_2 r_4 & 0 & 0 \\ B_1 \mu_1 & B_2 \mu_1 & B_3 \mu_1 & B_4 \mu_1 & -\mu_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\mu_2 & \mu_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\mu_2 & \mu_2 \\ B_1 \mu_2 & B_2 \mu_2 & B_3 \mu_2 & B_4 \mu_2 & 0 & 0 & 0 & -\mu_2 \end{pmatrix}$$

Stationary probability of the chain

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7, \pi_8)$$

is given by the balance equations  $\boldsymbol{\pi} \cdot \mathbf{Q}_0 = \mathbf{0}$  and normalisation condition  $\sum_{\forall i} \pi_i = 1$ :

Figure 13: Markov chain for  $\text{HM}_4\text{-HE}_{1,3}$  source

$$\begin{aligned} \pi_2 &= \frac{B_2 \cdot r_1}{B_1 \cdot r_2} \pi_1, & \pi_3 &= \frac{B_3 \cdot r_1}{B_1 \cdot r_3} \pi_1, & \pi_4 &= \frac{B_4 \cdot r_1}{B_1 \cdot r_4} \pi_1, \\ \pi_5 &= \frac{p_1 \cdot r_1 \cdot \mu_2}{B_1 \cdot p_2 \cdot \mu_1 (\mu_1 + \mu_2)} \pi_1 & \pi_6 = \pi_7 = \pi_8 &= \frac{r_1}{B_1 \cdot (\mu_1 + \mu_2)} \pi_1 \end{aligned}$$

$$\pi_1 = \frac{B_1 \cdot p_2 \cdot r_2 r_3 r_4 \cdot \mu_1 (\mu_1 + \mu_2)}{r_1 [3p_2 \cdot \mu_1 + p_1 \cdot r_2 r_3 r_4] + p_2 \cdot r_1 \cdot \mu_1 (\mu_1 + \mu_2) [B_4 \cdot r_2 r_3 + B_3 \cdot r_2 r_4 + B_2 \cdot r_3 r_4]}$$

$P_{\text{ON}} = \pi_1 + \pi_2 + \pi_3 + \pi_4$  and  $P_{\text{OFF}} = \pi_5 + \pi_6 + \pi_7 + \pi_8$  are probabilities that the source is in state On or Off. The probabilities for parameters of VoIP and Mean average packets  $\eta$  generated by source given above are

$\Delta = 20 \text{ ms}$ :

$$\begin{aligned} \boldsymbol{\pi} &= (0.00254, 0.02538, 0.16009, 0.18740, 0.56453, 0.02002, 0.02002, 0.02002) \\ P_{\text{ON}} &= 0.37541, & P_{\text{OFF}} &= 0.62459, & \eta &= \lambda \cdot P_{\text{ON}} = 18.8 \text{ p/s} \end{aligned}$$

$\Delta = 200 \text{ ms}$ :

$$\begin{aligned} \boldsymbol{\pi} &= (0.00018, 0.00290, 0.03615, 0.51164, 0.36371, 0.02847, 0.02847, 0.02847) \\ P_{\text{ON}} &= 0.55087, & P_{\text{OFF}} &= 0.44913, & \eta &= \lambda \cdot P_{\text{ON}} = 27.5 \text{ p/s} \end{aligned}$$

$\Delta = 1000 \text{ ms}$ :

$$\begin{aligned} \boldsymbol{\pi} &= (0.00007, 0.00135, 0.01889, 0.56227, 0.04151, 0.12530, 0.12530, 0.12530) \\ P_{\text{ON}} &= 0.58260, & P_{\text{OFF}} &= 0.41740, & \eta &= \lambda \cdot P_{\text{ON}} = 29.1 \text{ p/s} \end{aligned}$$

## 2 Token Bucket System

### 2.1 General steady-state analysis

As we have seen in the previous chapter, speech with applied VAD is a variable bit rate stream that generates burst traffic into the network. Using VAD, the same speech quality can be obtained with lower link capacity e.g 27.5 p/s for VAD with 200 ms fill-in time. This allows offering IP telephony on lower price, when it is agreed in SLA, and the customer is generating exactly this traffic. Anyhow if the customer violates SLA and he generates the worst-case traffic i.e. constant packet-rate 50 p/s, he uses link capacity 22,5 p/s free of charge. To avoid such as incorrect customer's behavior, provider has to use a policing mechanism, which inspects the incoming stream. Token bucket mechanism is very suitable policing mechanism for speech stream, because it introduces no additional delay and packets, which are out of SLA, are marked. We assume in the paper that the marked packets are discarded i.e. they are lost, what is the most rigid policing strategy.

Token Bucket System (TBS) is a buffer and a stream of tokens fills it continuously at a given rate. We assume a Poisson source of tokens. The maximum number of tokens in the queue is given by a token buffer depth "n", and when this queue is full, arriving tokens are rejected. If the speech packet is passing TBS, fixed number of tokens (due to fixed length of the speech packet) is removed from the token bucket queue. For better approximation of real TBS, and also for simplicity, we assume removing one token by one passing speech packet. If no token is available when the speech packet is passing TBS, the packet is marked and under our assumptions it is also rejected. Only unmarked packets will enter the network. Let  $P_{lst}$  is probabilities of packet marking, and we will call it "packet loss probability".

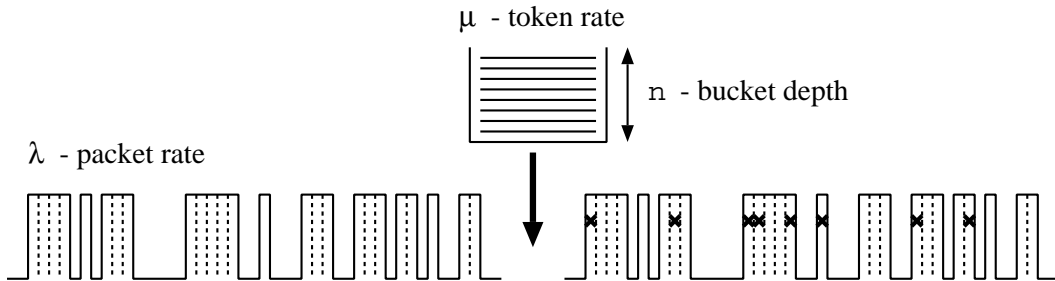


Figure 14: Token Bucket System

There are three elementary random phenomenons in TBS: packet arrival, token arrival and switching between On-Off states in the speech source. We assume that each of them is a Markov process, and therefore also TBS can be modelled by Markov process (see [4]). Fig.15 gives an idea of the transition diagram. Number of the column means the number of tokens in the token bucket. We have assigned this as "k-level" (see Fig.15).

Let On-Off Source has  $I$  On states and  $J$  Off states. Let  $R_k^i$  and  $S_k^j$  are probabilities that in TBS are  $k$  tokens and the Source is in On state  $i$ , eventually in Off state  $j$ . Let  $P_k$  is probability of  $k$ -level. Let variable  $q$  means the number of tokens in the token queue, and a state  $ON$  means that the speech source is in some of On states:

$$P_k = \sum_{i=1}^I R_k^i + \sum_{j=1}^J S_k^j \quad \text{and} \quad P_{lst} = P(q = 0 / ON) = \frac{P(q = 0 \ \& \ ON)}{P(ON)} = \frac{\sum_{i=1}^I R_0^i}{P_{ON}}$$

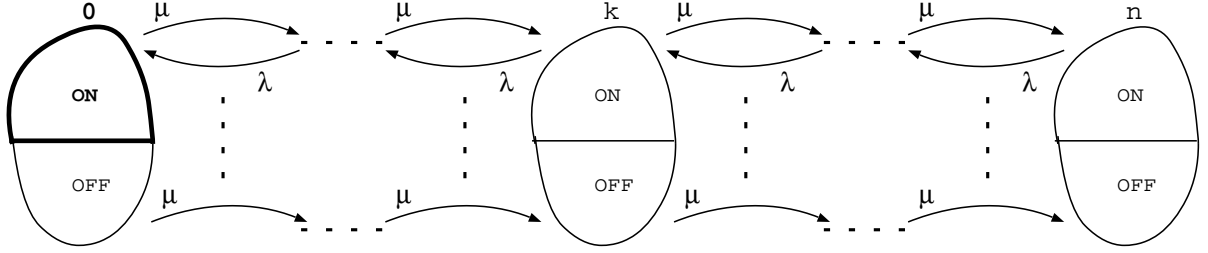


Figure 15: Markov chain of Token Bucket System

Taking from the Theory of Markov Chains we can solve balance equations  $\mathbf{p} \cdot \mathbf{Q} = \mathbf{0}$  for state probabilities of steady-state Markov chain, where  $\mathbf{p}$  is vector of state probabilities,  $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_n)$  where  $\mathbf{p}_k = (R_k^1, \dots, R_k^I, S_k^1, \dots, S_k^J)$  and  $\mathbf{Q}$  is the rate matrix.

We can write the rate matrix  $\mathbf{Q}$  to the block form:

$$\begin{pmatrix} \mathbf{A}_0 & \mathbf{\Lambda} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{\Gamma} & \mathbf{A} & \mathbf{\Lambda} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma} & \mathbf{A} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A} & \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Gamma} & \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{\Gamma} & \mathbf{A}_n \end{pmatrix}$$

The blocks  $\mathbf{A}_0$ ,  $\mathbf{A}$  and  $\mathbf{A}_n$  are matrixes of rates between states within 0-level,  $k$ -level (for  $k = 1, \dots, n - 1$ ) and  $n$ -level. The block  $\mathbf{\Gamma}$  is diagonal matrix with rate  $\lambda$  on first  $I$  positions,  $\mathbf{\Gamma} = \text{diag}\{\lambda, \dots, \lambda, 0, \dots, 0\}$  and  $\mathbf{\Lambda}$  is diagonal matrix with rate  $\mu$  on all  $I + J$  positions,  $\mathbf{\Lambda} = \text{diag}\{\mu, \dots, \mu\} = \mu \cdot \mathbf{E}$ . These all block matrixes have a size  $(I + J) \times (I + J)$ .

Let  $\mathbf{Q}_0$  is the rate matrix of On-Off Source. It's easy to see, that:

$$\mathbf{A}_0 = \mathbf{A} + \mathbf{\Gamma}, \quad \mathbf{A}_n = \mathbf{A} + \mathbf{\Lambda} \quad \text{and} \quad \mathbf{Q}_0 = \mathbf{\Gamma} + \mathbf{A} + \mathbf{\Lambda}$$

The balance equations  $\mathbf{p} \cdot \mathbf{Q} = \mathbf{0}$  have the following block form:

$$\begin{aligned} \mathbf{p}_0 \cdot \mathbf{A}_0 + \mathbf{p}_1 \cdot \mathbf{\Gamma} &= \mathbf{0} \\ \mathbf{p}_{k-1} \cdot \mathbf{\Lambda} + \mathbf{p}_k \cdot \mathbf{A} + \mathbf{p}_{k+1} \cdot \mathbf{\Gamma} &= \mathbf{0} \quad k = 1, \dots, n - 1 \\ \mathbf{p}_{n-1} \cdot \mathbf{\Lambda} + \mathbf{p}_n \cdot \mathbf{A}_n &= \mathbf{0} \end{aligned}$$

## 2.2 TBS for speech with noVAD

The previous section has introduced a general approach to the TBS modelling. Each TBS model depends on the chosen speech source model. Let's start with the simplest case, if no VAD is applied to speech, and packets transmit all samples of speech and silence (background noise). This speech packet traffic will be modelled by Poisson stream with rate  $\lambda = 50$  p/s, and the general state transition diagram given by Fig.15 will be reduced to the diagram that is shown on Fig.16. (loops are omitted).

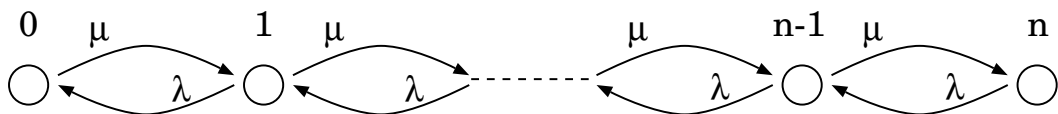


Figure 16: Markov model of TBS with no VAD speech traffic

The block form of the matrix shown above for the general source, will be given in this case by blocks:

$$\mathbf{\Gamma} = (\lambda), \quad \mathbf{\Lambda} = (\mu), \quad \mathbf{A}_0 = (-\mu), \quad \mathbf{A}_n = (-\lambda), \quad \mathbf{A} = (-(\mu + \lambda))$$

Let  $P_k$  is probability of  $k$  tokens in the token queue. Then the packet loss probability is given by probability of empty token queue:  $P_{lst} = P_0$ . Let's use symbol  $\rho = \mu/\lambda$  for traffic intensity. Equations describing steady state probability distribution are as follow:

$$\begin{aligned} 0 &= -\rho P_0 + P_1 \\ 0 &= \rho P_{k-1} - (\rho + 1)P_k + P_{k+1} \quad k = 1, 2, \dots, n-1 \\ 0 &= \rho P_{n-1} - P_n \end{aligned}$$

These well known equations give an Erlang distribution:

$$P_k = \rho^k P_0 \quad k = 1, \dots, n, \quad \sum_{k=0}^n P_k = 1 \quad \Rightarrow \quad P_0 = \left[ \sum_{k=0}^n \rho^k \right]^{-1}$$

and the packet loss probability when the bucket depth is  $n$ , is given by:

$$P_{lst}(n) = P_0 = \frac{\rho - 1}{\rho^{n+1} - 1}$$

### 2.3 TBS for speech with full VAD

Applying the HM<sub>2</sub>-HM<sub>2</sub> model of the speech source if all gaps are detected and removed (see Fig. 4) to the general TBS model (see Fig. 15), the transition diagram of the TBS system will be as follow (loops are omitted again):

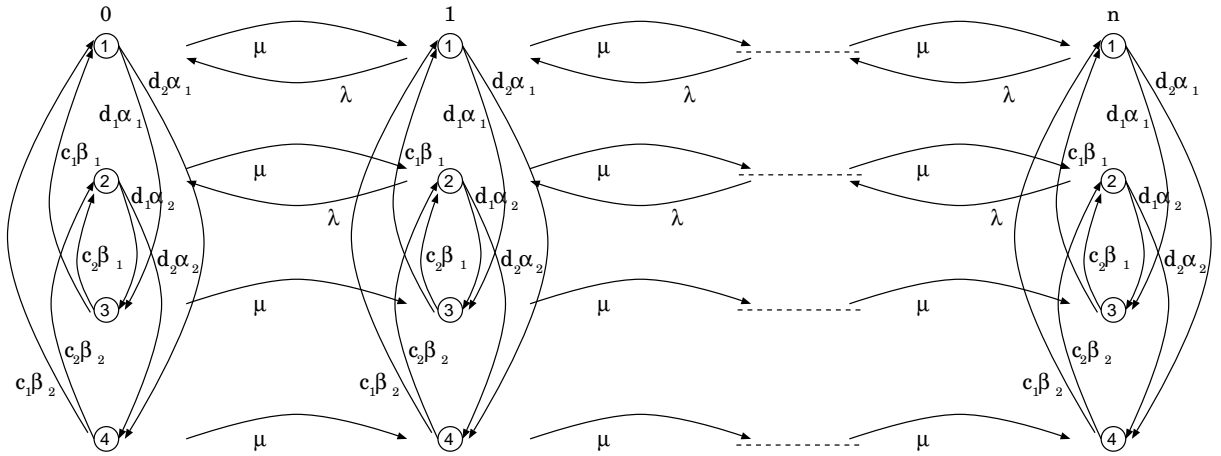


Figure 17: Markov chain of TBS for speech with full VAD

The blocks in the general rate matrix  $\mathbf{Q}$  mentioned above are:

$$\begin{aligned} \mathbf{A}_0 &= \begin{pmatrix} -(\mu+\alpha_1) & 0 & d_1\alpha_1 & d_2\alpha_1 \\ 0 & -(\mu+\alpha_2) & d_1\alpha_2 & d_2\alpha_2 \\ c_1\beta_1 & c_2\beta_1 & -(\mu+\beta_1) & 0 \\ c_1\beta_2 & c_2\beta_2 & 0 & -(\mu+\beta_2) \end{pmatrix} & \mathbf{A}_n &= \begin{pmatrix} -(\alpha_1+\lambda) & 0 & d_1\alpha_1 & d_2\alpha_1 \\ 0 & -(\alpha_2+\lambda) & d_1\alpha_2 & d_2\alpha_2 \\ c_1\beta_1 & c_2\beta_1 & -\beta_1 & 0 \\ c_1\beta_2 & c_2\beta_2 & 0 & -\beta_2 \end{pmatrix} \\ \mathbf{A} &= \begin{pmatrix} -(\mu+\alpha_1+\lambda) & 0 & d_1\alpha_1 & d_2\alpha_1 \\ 0 & -(\mu+\alpha_2+\lambda) & d_1\alpha_2 & d_2\alpha_2 \\ c_1\beta_1 & c_2\beta_1 & -(\mu+\beta_1) & 0 \\ c_1\beta_2 & c_2\beta_2 & 0 & -(\mu+\beta_2) \end{pmatrix} & \mathbf{\Gamma} &= \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} & \mathbf{\Lambda} &= \begin{pmatrix} \mu & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & \mu \end{pmatrix} \end{aligned}$$

## 2.4 TBS for speech with partial VAD

Applying the HM<sub>4</sub>-HE<sub>1,3</sub> model of the speech source if VAD with positive finite fill-in time is used (see Fig. 12) to the general TBS model (see Fig. 15), the transition diagram of the TBS system will be as follow (loops are omitted again):

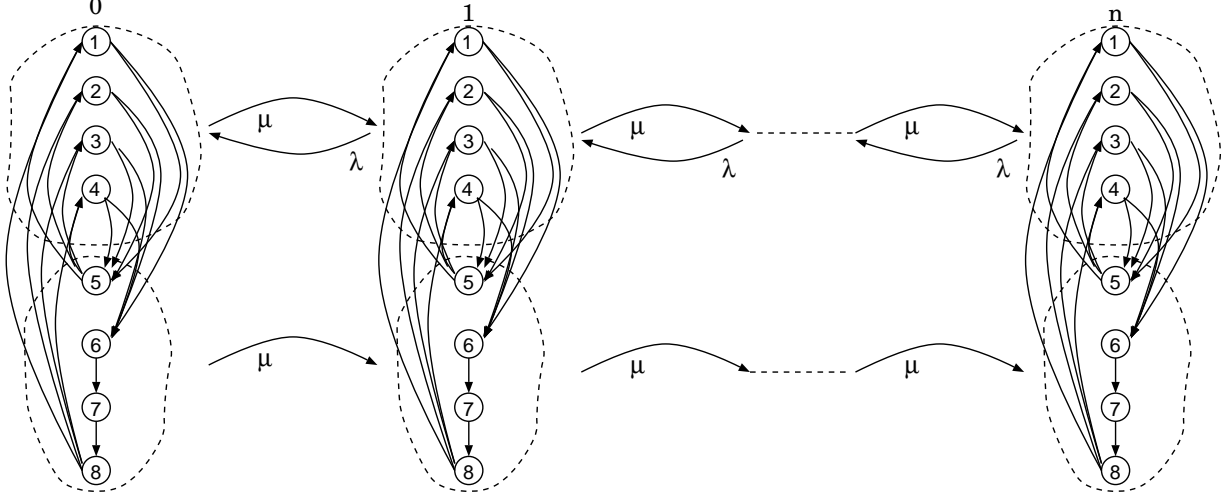


Figure 18: Markov chain of TBS for speech with partial VAD

The blocks in the general rate matrix  $\mathbf{Q}$  mentioned above are:

$$\mathbf{A}_0 = \begin{pmatrix} -(\mu+r_1) & 0 & 0 & 0 & p_1 r_1 & p_2 r_1 & 0 & 0 \\ 0 & -(\mu+r_2) & 0 & 0 & p_1 r_2 & p_2 r_2 & 0 & 0 \\ 0 & 0 & -(\mu+r_3) & 0 & p_1 r_3 & p_2 r_3 & 0 & 0 \\ 0 & 0 & 0 & -(\mu+r_4) & p_1 r_4 & p_2 r_4 & 0 & 0 \\ B_1 \mu_1 & B_2 \mu_1 & B_3 \mu_1 & B_4 \mu_1 & -(\mu+\mu_1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -(\mu+\mu_2) & \mu_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -(\mu+\mu_2) & \mu_2 \\ B_1 \mu_2 & B_2 \mu_2 & B_3 \mu_2 & B_4 \mu_2 & 0 & 0 & 0 & -(\mu+\mu_2) \end{pmatrix}$$

$$\mathbf{A}_n = \begin{pmatrix} -(r_1+\lambda) & 0 & 0 & 0 & p_1 r_1 & p_2 r_1 & 0 & 0 \\ 0 & -(r_2+\lambda) & 0 & 0 & p_1 r_2 & p_2 r_2 & 0 & 0 \\ 0 & 0 & -(r_3+\lambda) & 0 & p_1 r_3 & p_2 r_3 & 0 & 0 \\ 0 & 0 & 0 & -(r_4+\lambda) & p_1 r_4 & p_2 r_4 & 0 & 0 \\ B_1 \mu_1 & B_2 \mu_1 & B_3 \mu_1 & B_4 \mu_1 & -\mu_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\mu_2 & \mu_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\mu_2 & \mu_2 \\ B_1 \mu_2 & B_2 \mu_2 & B_3 \mu_2 & B_4 \mu_2 & 0 & 0 & 0 & -\mu_2 \end{pmatrix}$$

$$\mathbf{A}_0 = \begin{pmatrix} -(\mu+r_1+\lambda) & 0 & 0 & 0 & p_1 r_1 & p_2 r_1 & 0 & 0 \\ 0 & -(\mu+r_2+\lambda) & 0 & 0 & p_1 r_2 & p_2 r_2 & 0 & 0 \\ 0 & 0 & -(\mu+r_3+\lambda) & 0 & p_1 r_3 & p_2 r_3 & 0 & 0 \\ 0 & 0 & 0 & -(\mu+r_4+\lambda) & p_1 r_4 & p_2 r_4 & 0 & 0 \\ B_1 \mu_1 & B_2 \mu_1 & B_3 \mu_1 & B_4 \mu_1 & -(\mu+\mu_1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -(\mu+\mu_2) & \mu_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -(\mu+\mu_2) & \mu_2 \\ B_1 \mu_2 & B_2 \mu_2 & B_3 \mu_2 & B_4 \mu_2 & 0 & 0 & 0 & -(\mu+\mu_2) \end{pmatrix}$$

$$\mathbf{\Gamma} = \begin{pmatrix} \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda \end{pmatrix}$$

$$\mathbf{\Lambda} = \begin{pmatrix} \mu & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}$$

## 3 Impact of TBS to speech quality

### 3.1 Speech quality

Due to random character of spurt and gap duration in speech at the input of TBS, we have to accept some speech quality degradation. As we have mentioned in Introduction, the only quality degradation factor caused by TBS is the packet loss. Our approach to evaluation of the packet loss impact to the speech quality is based on the E-model [5]. The E-model was modified according [6], [7]. The reason is that in the standardised E-model, the impact of the packet loss is an inseparable part of the codec degradation factor, and the model does not allow an independent setting of the VAD fill-in time and the packet loss. On the other side, the used modification takes impact of the packet loss as an impairment, which occurs simultaneously with the speech. This approach allows applying all VAD modification to speech, and afterwards evaluating an impact of the packet loss to the speech quality. To underline the impact of fill-in time, token rate and token queue depth to the speech quality, configuration parameters were set as simple as possible: codec G.711 with 20 ms of speech in one packet, zero delay within network (look-ahead delay in fill-in based VAD was not taken into account), and all other parameters of E-model were set to default values according [5].

### 3.2 Impact of TBS parameters to speech quality

The following configuration is assumed for the study: - speech is generated by G.711 codec and one packet is filled by 20 ms frame of speech - there are spurt and gap periods in the speech, and they are distributed according [1] - packet flow is generated by Markov models described in the previous sections - Voice Activity Detector bridges over gaps shorter than given fill-in time without look-ahead delay - token flow is creating by Poisson source - speech packets, passing TBS when no token is available, are rejected - modified E-model mentioned in the previous section is used for QoS evaluation Behaviour of this configuration was under study and the following values of parameters were selected. VAD fill-in time was set to zero (if an "ideal" VAD is applied), 20 ms (fill-in time is equal to one packet speech segment), 200 ms (typical setting for traditional VADs), 1 s (very high value), and infinity (no VAD is applied). A set of token bucket depth values includes 10, 20, 30, 40, 70, and token rate growths from 10 token/s to 60 token/s by 5 token/s increment.

First of all we have to say, that we are not familiar with any tool allowing evaluation of VAD fill-in time to speech quality. E-model does not take into account impact of clipping, replacing background noise by comfort noise, or impact of short gaps to the adaptive jitter buffer. Then we assume, that VAD manufacturer will select the appropriate fill-in time. According [3], spurt and gaps distribution in speech leaving G.729 Annex B VAD is very close to the distribution produced by traditional silence detector without fill-in. Then zero fill-in time will represents VAD according mentioned recommendation despite of no compression assumed in the model. The aim of the paper is to show impact of token bucket parameters (token rate, token bucket depth) to voice quality.

We can see graphical representations of these results on the following figures.

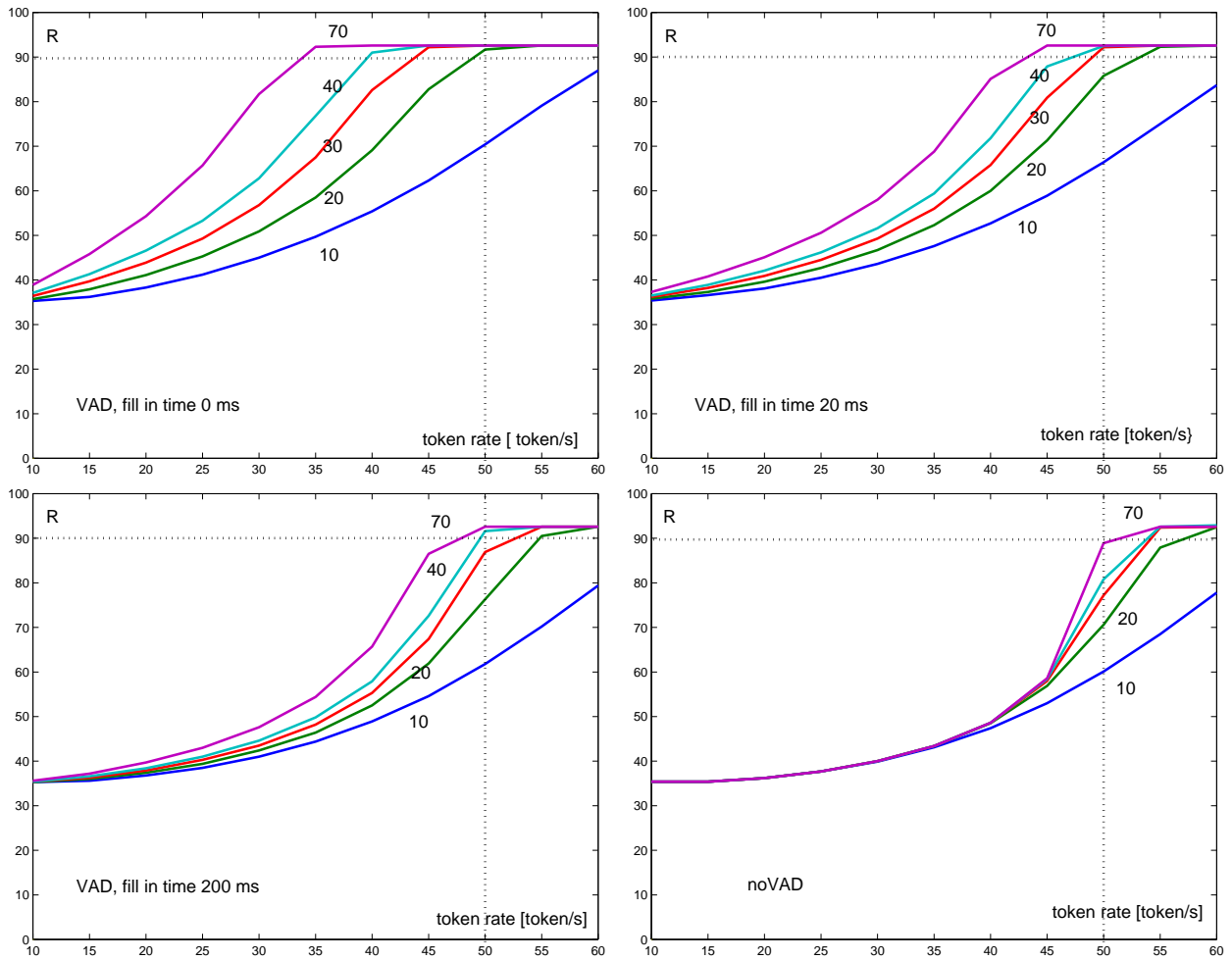


Figure 19: Voice quality as a function of TBS queue depth  $n$  and token rate  $\mu$

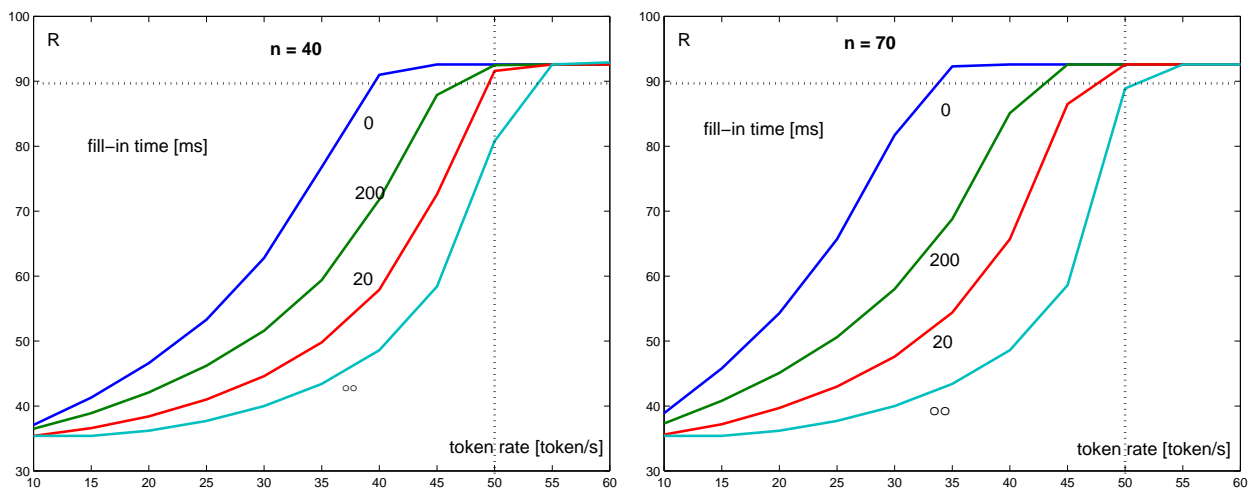


Figure 20: Voice quality as a function of VAD fill-in time and token rate  $\mu$

Because no degradation factors of the network is taken into account, we assume, that TBS parameters will be set to keep QoS over the limit  $R=90$ . The figures allow for given fill-in time to set appropriate values of TBS parameters. As an example, the following figure shows values of TBS queue depth and token rate giving exactly  $R=90$  limit assuming fill-in time 0 ms i.e. full VAD or G.729 Annex B VAD.

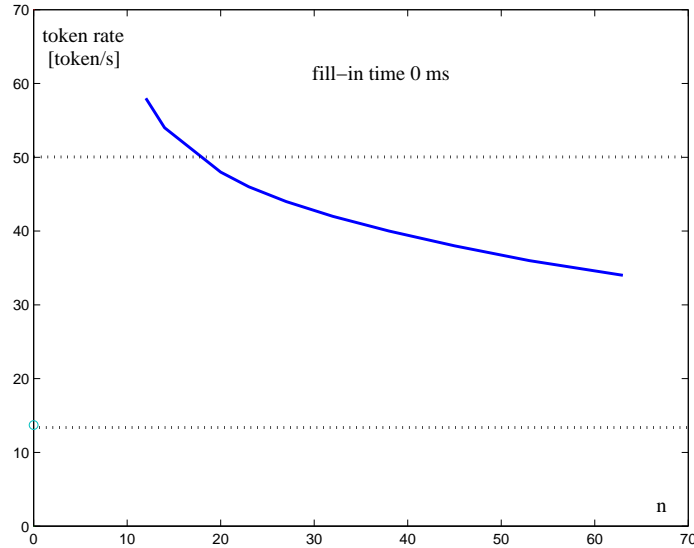


Figure 21: Token rate as a function of TBS queue depth for voice quality  $R = 90$

The lower limit of the packet rate entering the network is given by an average packet rate at the VAD output depending on the chosen fill-in time. Figure 22 shows, that any higher link capacity savings by additional decreasing of the token rate leads to packet loss increasing, and consequently to the rapid degradation of speech quality (see Fig. 23).

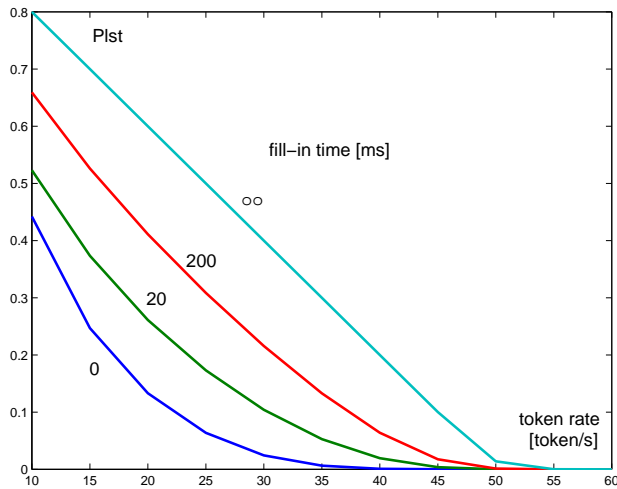


Figure 22: Packet loss probability as a function of token rate

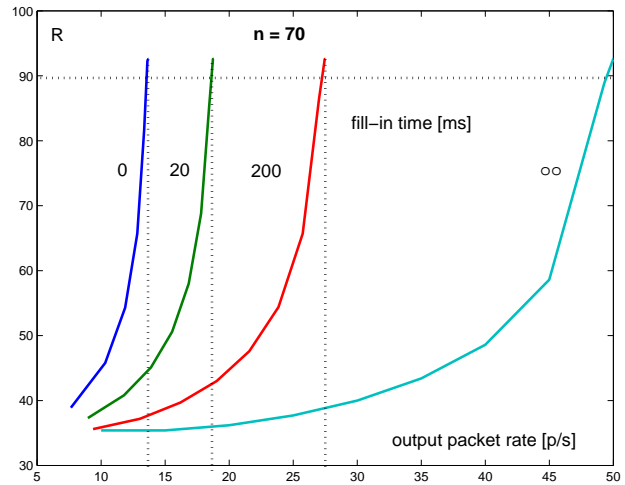


Figure 23: Voice quality as a function of output packet rate

fill in time	$\eta$	mean spurt	mean gap
0 ms	13.7 p/s	227 ms	596 ms
20 ms	18.8 p/s	511 ms	798 ms
200 ms	27.5 p/s	3002 ms	2453 ms
1000 ms	29.1 p/s	4730 ms	3381 ms

## 4 Conclusions

Voice Activity Detector is the main tool that can be used to save link capacity allocated to the telephone call in the packet switched network. VAD application leads to changing of the speech traffic shape, which is modified from constant bit-rate to the burst traffic. To avoid additional delay, the peak rate has to be maintained during a talk-spurt, and capacity saving is obtained by silence suppression. This speech source can be modelled by Markov chain that is switched between two subsets of states: On (spurt) and Off (gap), and time interval that is needed to pass one of the chain subsets has phase distribution. During On period the source emits packets as a Poisson process as an approximation of a regular emitting in reality. On the other side, the network has to be protected from incorrect behaviour of the customer, which can generate higher bit-rate (up to peak-rate) than it is agreed in SLA. Therefore some policing mechanism at the input to the network is needed. Token bucket system seems to be a suitable policing for speech, anyhow it may cause some packet loss due to random behaviour of spurt/gap intervals. The paper gives a method how to set parameters (token rate and token queue depth of TBS) to preserve high speech quality at the defined level. It may be interesting to find the worst case behaviour of the customer against TBS with some speech QoS guarantee. Unfortunately this problem is out of scope of this paper.

## References

- [1] ITU-T Recommendation P.59, "*Artificial Conversational Speech*", 1993
- [2] J. G. Gruber, "*A Comparison of Measured and Calculated Speech Temporal Parameters Relevant To Speech Activity Detection*", IEEE Transactions on Communications, Vol. COM-30, No. 4, pp. 728–738, April 1982.
- [3] W. Jiang and H. Schulzrinne. "*Analysis of on-off patterns in voip and their effect on voice traffic aggregation*". In The 9th IEEE International Conference on Computer Communication Networks, 2000.
- [4] J. Smieško, "*Základy teórie hromadnej obsluhy (Queueing Theory)*", textbook, pp. 190, MC Energy Žilina, ISBN 80-968115-6-8, (1999)
- [5] ITU-T Recommendation G.107, "*The E-model, a computational model for use in transmission planning*", 2003
- [6] M. Klimo, "*Cell Loss Noise in the Case of Linear Reconstruction*", International Conference on Telecommunications, Proceedings of the ICT'98, Porto Carras, Greece, 21-25 June 1998,
- [7] M. Klimo, "*Voice over IP: Packet Loss and Jitter in E-model*", Proceedings of the QoS Summit'99, Paris, France, November 1999