Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol 2, No.5, 2012

Estimating the Quality of Digitally Transmitted Speech over Satellite Communication Channels

Aderemi A. Atayero* Adeyemi A. Alatishe Juliet O. Iruemi

Department of Electrical and Information Engineering, Covenant University, PMB 1023, Ota, Ogun State, Nigeria

* E-mail of the corresponding author: atayero@covenantuniversity.edu.ng

Abstract

Analogue speech signal is one of the most natural means used by humans for communication purposes. The emergence of digital modulation and coding techniques has made the transmission of analogue speech (as digital content) over various conduits possible, albeit with inevitable signal degradation as a result of errors inherent in the conversion process. A need naturally arises for determining the quality of speech received at the information sink, with a view to enhancing its robustness to degradation suffered in transit over the communication channel. We present in this paper analytic methods of qualitative assessment of the quality of recovered digitally transmitted speech. A methodology for determining the intelligibility of speech by using segmental SNR gotten by dividing the speech signal into M integer segments is proposed. This methodology has the following advantages: a) it allows for assessing the dynamics of change of speech quality in real-time through statistical modeling, b) it obviates the need for expensive, yet subjective experimental approaches like MOS, and c) it takes into consideration not only the signal power, but also its spectral characteristics which is a step above the use of Modulated Noise Reference Units (MNRUs). Using the obtained results, a procedure for analysis of speech intelligibility by means of statistical modeling is developed.

Keywords: Speech processing, Mean opinion score, MOS, SNR, PCM, Quantization noise

1. Introduction

The criteria and methods of estimating the quality of speech reproduction and recovery are classified into two major groups; objective and subjective. The objective group employs certain formalized parameters, capable of determining the degree of divergence between the original and reproduced speech. Humans serve as the information sink and as such the most important element of any telecommunication system; hence signal quality is assessed subjectively by our perception of transmitted speech. It is common practice to employ procedures using the Mean Opinion Score (MOS) of groups of experts (ITU-T P.800, 1996a; ITU-T P.800.1, 1996b; ITU-T P.830, 1996c) in assessing the quality of speech channels. In which case, the quality of perception of transmitted speech signal is measured using a 5-scale system as presented in Table 2. Processing the scores given by groups of expert listeners after listening to various speech signals played back through different loud speakers gives the MOS estimates. Each listener gives a score for each of the signals using the scaling in Table 1, the results are then averaged. Figure 1 shows the MOS score for various coding methods (Atayero, 2000). While signal quality has a direct correlation with transmission speed, more complex algorithms are capable of achieving higher quality to transmission speed ratio.

In line with the criteria for accurate reproduction of speech signal given in (Atayero, 2000; Bishnu and Schroeder 1979), it is possible to isolate the indicator of accurate reproduction of both individual realizations of the signal as well as of groups of realization. Mean-square approximation indicators are generally preferred. The subjective criteria of estimating quality of digitally transmitted speech are used for measurements involving experts. Subjective quality indicators are determined via the direct use of the human auditory organs. The articulate method intelligibility criterion is the most popularly adopted. This method is based on measuring the intelligibility S% of received speech, which is defined by the percentage of correctly received speech elements like; sounds, syllables, words, or phrases. Under certain types of distortion, intelligibility is functionally linked to other quality measures e.g. Signal-to-Noise Ratio (SNR), and it adequately characterizes quality.

Occurrence of error bits in the transmission of speech over digital satellite communication channels worsens the quality of signal recovery, and consequently the intelligibility of recovered speech signal. Analysis of intelligibility

of such systems is tied to the problem of estimating the power of the additional noise caused by the loss of speech bits. This in essence is the estimation of the discretization and recovery noise under random change of discretization frequency conditions. These in conjunction with quantization noise present in digital communication systems together with additive noise determine the quality of speech perception, which is most often estimated as syllabic intelligibility – S% (Atayero, 2000).

The normalized error indicator is often used for quantitative estimation of the quality of speech signal reception. It characterizes the mean square error (MSE) of reception σ^2_{N}, averaged in time and normalized with information variance σ^2_{λ}:

$$\bar{\delta}^2 = \frac{\sigma^2_N}{\sigma^2_\lambda} \tag{1}$$

where σ^2_N - Noise variance.

The inverse error quantity is the ratio of signal power to noise power.

$$SNR = 10log\left(\bar{\delta}^2\right) dB \tag{2}$$

Thus, for the analysis of any speech transmission system, it is necessary to estimate the ratio of signal power to the total noise power, denoted as SNR_{Σ}, and determine the correlation between SNR_{Σ} and S%. When considering the transmission of speech signal over analogue channels, the decibel value of the SNR is often used for characterizing the transmission conditions.

$$A = 10log(D_s/D_e) \tag{3}$$

The SNR values have a stable correlation with the subjective estimates of the quality of speech perception. The numerical characteristics of intelligibility of speech fragments (phonemes in particular) is majorly used as metrics of subjective estimates.

A correlation function for syllabic intelligibility {S*} with other forms of intelligibility: word, phrase, phoneme has been established (Atayero, 2000). Since expression (3) employs both signal (Ds) and noise (De) variance calculated (or measured) for the whole test duration of the speech signal, this indicator is called the long-term SNR.

Suffice it to mention here that research into digital methods of speech transmission and specifically different adaptive methods of modulation has shown serious discrepancies in subjective estimates of same values of A (Kitawaki, Honda, and Itoh, 1984). This can be attributed to the varying nature of distortion caused by both adaptive and non-adaptive transmission systems. In the latter case, we have the presence of stationary noise whose level is independent of the signal level. The quality of communication channel in this case is determined majorly via the perception of noise level during pauses in speech transmission.

The noise of unoccupied channels may be undetectable to the ear in adaptive systems. In this case, the perception of distortion in reproduced speech will be determined by accompanying non-stationary noise, the variance of which is determined by both the signal level and its spectral characteristics. In connection with this, for the subjective estimation of different algorithms of coding and recovering speech, special devices are employed for generating noise in correlation with the speech signal. Such devices are called Modulated Noise Reference Unit (MNRU) (Perkins et al. 1997). The use of MNRU allows for taking into consideration the non-stationarity of noise occurring as a result of changes in the instantaneous power of speech signal. We note here however, that change in signal spectral model during the pronunciation of vocalized and non-vocalized sounds is not taken into consideration.

On the other hand, some works reported in the literature have shown that stable statistical correlation between objective and subjective estimates in the analysis of speech transmission systems with adaptive modulation methods under different algorithms can be achieved if the quantity in expression (4) is adopted as the objective estimate:

$$A_S^* = \frac{1}{M}\sum_{k=1}^{M} SNR(k) \tag{4}$$

where $SNR(k) = 10\log\frac{D_S(k)}{D_\varepsilon(k)}$ – ratio of signal power to noise power, computed for the kth time window of the speech signal, containing N measurements; M – number of sequential speech test-signal windows, for which $SNR(K)$ is averaged.

We consider A_S^* - as value of segmental SNR. Note that A_S^* can be estimated for fragments of speech signal as well as for whole speech tests. The M value should chosen taking into consideration the objective of the task at hand. sizing.

2. Speech overload and quantization noise power

Assuming that the signal is evenly distributed across quantization steps, then quantization noise equals $\lambda^2/12$. Let $\pm L$ represent the limit of change in amplitude of the input signal and $W(\lambda)$ be the probability flux density of instantaneous values of input signal.

$$P_{Tx} = \int_{L\Delta}^{\infty}\left(\frac{1}{2}L\Delta - \lambda\right)^2 \cdot W(\lambda)d\lambda \tag{5}$$

For a majority of practical cases, the overload level is usually taken as equal $+3$ dB. The overload noise power is easily calculated from the pfd models of speech signal. We note here that quantization and limiting noise do not occur simultaneously (since each corresponds to different samples of the signal, which are weakly correlated for standard digital transmission system). Therefore the total noise power occurring in the process of quantization is the sum of these two components.

The use of linear quantization for the transmission of telephone signals is not optimal for the following reasons: the amplitude distribution of analogue speech signal is not uniform, low signal amplitudes are more probable than their high counterparts. In which case an increase in the quantization SNR if quantization error for more probable amplitudes be reduced comes as a given.

Analogue speech signal can change by up to $60\ dB$, for this reason, it is not easy to achieve with a regular low level signal quantization codec the same accuracy as for those of higher level. Optimization of the compression function for noise minimization can only be carried out for a specific signal with known statistical characteristics. A deviation from the a priori parameters of the signal results in a significant increase in quantization noise power. Non-uniform quantization (coarse quantization process of high-level signals and the precise quantization of low-level signals) is used in real systems with digital PCM. This is achieved through the use of a compressor at the receiving end. In practice, modifications of the logarithmic function of compressor is employed:

A - characteristic (Jayant and Noll, 1984).

$$y = \begin{cases} \frac{A\cdot|\lambda|}{1+\ln A} & , \quad 0 \le |\lambda| \le A^{-1} \\ \frac{1+\ln(A\cdot|\lambda|)}{1+\ln A} & , \quad A^{-1} \le |\lambda| \le 1 \end{cases} \tag{6}$$

and μ - characteristic

$$y = \frac{\ln(1+\mu\lambda)}{\ln(1+\mu)} \; ; \quad 0 \le |\lambda| \le 1 \tag{7}$$

It has been established that with sufficient quantization level L, the quantization noise power depends not only on compression characteristics (6) or (7), but also on the probability flux density of instantaneous values of the speech signal.

$$\sigma_q^2 = \frac{1}{3L^2}\int_{-1}^{1} W(\lambda)\cdot\left[y'^{(\lambda)}\right]^{-2} d\lambda \tag{8}$$

Hence, the quantization SNR {SNRq} is defined as

$$SNR_q = \frac{\sigma_\lambda^2}{\sigma_q^2} = 3L^2\cdot\sigma_\lambda^2\left\{\int_{-1}^{1}\frac{W(\lambda)}{\left[y'^{(\lambda)}\right]^2}\cdot d\lambda\right\}^{-1} \tag{9}$$

When the μ compression characteristic is employed, logarithmic quantization is used for all quantization levels of the speech signal. Then from (7) and (8) we obtain:

$$\sigma_q^2 = \frac{\ln^2(1+\mu)}{3L^2}\int_{-1}^{1} W(\lambda)\cdot\left[\frac{1}{\mu}+\lambda\right]^2 d\lambda \tag{10}$$

And SNR_q will be of the form

$$SNR_q = \frac{3L^2}{\ln^2(1+\mu)}\left[1+\frac{1}{\mu^2\cdot\sigma_\lambda^2}+\frac{\int_{-1}^{1}W(\lambda)\cdot d\lambda}{\mu\cdot\sigma_\lambda}\right]^{-1} \tag{11}$$

In line with expression (7), for the estimation of average quantization noise σ_q^2 the expression for averaged variance of quantization error is widely used.

$$\sigma_q^2 = \frac{\Delta^2}{6}\cdot\int_0^{\infty}\omega(\lambda)\cdot\left[y'^{(\lambda)}\right]^{-2} d\lambda \tag{12}$$

Inserting the pfd of speech signal $\omega(\lambda)$ in expression (12) and one of the quantization characteristics (7) or (12), the quantization noise power can be estimated. Using (2) and (12) we arrive at the quantization SNR.

$$\mathrm{SNR_q} = 10\log\left(\frac{\sigma_q^2}{\sigma_\lambda^2}\right)^{-1} \tag{13}$$

3. Discretization and Recovery Noise variance

Real continuous signal presented as sampled data at the input of an interpolation filter, will be recovered with a given amount of interpolation error. It is a fact established that for a linear system, the discretization error is made up of two component: a) dynamic component that occurs as a result of the distortion of useful message when passing through the interpolating device and b) interference component, which appears as a result of spectrum offset components of discrete samples falling within the bandwidth of the interpolating device (Milner and Semnani 2000). As a result, the variance of total error can be calculated from the expression:

$$\sigma_{\Sigma_\vartheta}^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty}|W_0(j\omega) - W_P(j\omega)|^2 S_\lambda(\omega)\,d\omega + \frac{1}{2\pi}\int_{-\infty}^{+\infty}|W_P(j\omega)|^2 S_{CM}(\omega)\,d\omega \;= \sigma_\vartheta^2 + \sigma_i^2 \tag{14}$$

where σ_ϑ^2 ; σ_i^2 – dynamic and interference component variance respectively; $W_0(j\omega)$ – complex transfer coefficient of an ideal interpolator; $S_\lambda(\omega)$ – psd of the dynamic error component; $W_P(j\omega)$ – complex transfer coefficient of a real interpolator;

$$S_{CM}(\omega) = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} S_\lambda\left(\omega - \frac{\omega k}{T}\right) \tag{15}$$

where $S_{CM}(\omega)$ – *psd* of interference component of error.

Similar to the above stated, the discretization SNR can be obtained as given in (16)

$$SNR_\vartheta = 10\log\left(\frac{\sigma_{\Sigma_\vartheta}^2}{\sigma_\lambda^2}\right)^{-1} \tag{16}$$

As an illustration of the expressions given above, we consider the case of an interpolating device with the transfer function given in equation (17)

$$|W_P(j\omega)|^2 = \left[1 + \left(\frac{\omega}{2\pi F_{av}}\right)^{2m}\right]^{-1} \tag{17}$$

$$F_{av} = 3.4\ kHz, F_0 = 400\ Hz, \alpha = 1; 1.5; 2; 2.5\ kHz$$

The speech signal *psd* model is as given (18)

$$S_\lambda(\omega) = 2\alpha\left[\frac{1}{\alpha^2 + 4\pi^2\cdot(f+f_0)^2} + \frac{1}{\alpha^2 + 4\pi^2\cdot(f-f_0)^2}\right] \tag{18}$$

The functional relationship of the transfer function and speech information *psd* is presented in Fig. 3 with the following respective labels:

1–filter transfer function for $m = 1$; 1–filter transfer function for $m = 2$; 3–filter transfer function for $m = 3$;

4–filter transfer function for $m = 10$; 5–normalized spectrum $S_1(\omega)$.

The relationships presented in the figures allow for the qualitative estimation of recovery error, while varying the characteristics of the interpolator and speech signal *psd* parameters. Figures 3a and 3b depict the transfer characteristics of the interpolation filter as well as the relationships of the offset spectra $S_\lambda^*(\omega) = \sum_{k=1}^{2} S_\lambda(\omega - k\Omega_\partial)$, which allows for determining the source and magnitude of interference component of recovery error.

4. Communication Channel SNR

In addition to the above mentioned factors affecting speech intelligibility during transmission over satellite communication channels, like any other digital communication system, the transmission quality is also estimated via channel signal-to-noise ratio. The comunication channel SNR (SNR_{cc}) is defined by error bit of an element of digital signal in the communication channel (19).

$$P_{err} = f(SNR_{cc}) \tag{19}$$

In the presence of WGN in the communication channel, $SNR_{cc} = E/N_0$, where E is energy of the transmitted signal;

N_0 is the spectral density of additive white noise.

For the transmission of binary symbols at a rate $R = (T_0 l)^{-1}$, where T_0 is the discretization interval length; l –average number of bits in information symbol η_{liq}, in a channel with bandwidth *B,* the lower bound on probability of error for amplitude modulation (AM), frequency modulation (FM) and phase modulation (PM) and coherent detection satisfies the inequality given in (20).

$$P_{err} < \exp\left(-\frac{P_s/N_0 B}{2\bar{l}}\right) \tag{20}$$

where P_s – signal power.

For a more accurate estimate of the function $P_{err} = f(SNR_{cc})$, it is necessary to determine the modulation type, frequency characteristics of the channel $K(\omega)$ as well as the mode of reception. For a channel with Gaussian noise error probability distribution under optimal reception of binary symbols for FM and PM, equation (19) becomes:

$$P_{err} = 0.5 - \Phi\left(SNR_{cc} \cdot \sqrt{1 - r}\right) \tag{21}$$

where $r = -1$ for PM; $r = 0$ for FM; $\Phi(\cdot)$ – probability integral.

In a channel with inter-symbol interference, the error probability will increase due to the prevailing tendency of error grouping. However, if the receive signal is subjected to optimal nonlinear processing on the basis of a Viterbi processor, then P_{err} can be defined by Forney's ratio (Forney, 1972).

$$\frac{K_l}{2}\Phi\left(\frac{d_{min}^2}{2\sigma_\xi}\right) \le P_{err} \le \frac{K_u}{2}\Phi\left(\frac{d_{min}^2}{2\sigma_\xi}\right) \quad (22)$$

where K_l and K_u are constant coefficients; σ_ξ –additive noise variance in signal bandwidth; d_{min}^2 – energy of received signal in the presence of error. We note that the error probability in this case differs only slightly from the boundary value (20). The communication channel SNR can be gotten by specifying one of (19), (20), (21) in the form:

$$SNR_{cc} = f^{-1}(P_{err}) \quad (23)$$

5. Conclusion

Generally, the sink of digitally transmitted speech is the human auditory system. This has informed the most popular means of estimating quality of digitally transmitted speech i.e. MOS, which is based on the subjective perception of quality by a group of experts. The decibel value of Signal-to-Noise Ratio (SNR) was used in characterizing the process of speech transmission over analogue channels. Assessments of various digital transmission methods, especially different adaptive methods of modulation show substantial discrepancy between subjective assessments of speech (e.g. using Mean Opinion Score MOS) under similar SNR conditions was conducted. The segmental approach to determining the SNR of received speech as objective measure of quality is adopted. Analytic estimation of overload and quantization noise power, discretization and recovery noise variance, as well as the SNR of the communication channel as components of the total SNR_Σ are presented.

References

Atayero A. (2000), "Estimation of the Quality of Digitally Transmitted Analogue Signals over Corporate VSAT Networks", *PhD Thesis*, MTUCA.

Bishnu S.A., Schroeder M. R. (1979), "Predictive coding of speech signals and subjective error criteria. IEEE Transactions on Acoustics, Speech and Signal Processing, pages 247--254, June 1979.

Forney G. Jr. (1972), "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *Information Theory, IEEE Transactions on*, vol.18, no.3, pp. 363- 378, May 1972.

ITU-T P.800 (1996a), "Recommendation P.800 of the International Telecommunication Union, Methods for subjective determination of transmission quality", ITU-T, 1996.

ITU-T P.800.1 (1996b), "Mean Opinion Score (MOS) terminology", ITU-T, July 2006.

ITU-T P.830 (1996c), "International Telecommunication Union Recommendation, Subjective performance assessment of telephone-band and wideband digital codecs", February 1996.

Jayant, N. and P. Noll (1984), "Digital Coding of Waveforms—Principle and Applications to Speech and Video Englewood Cliffs", New Jersey: Prentice-Hall, 1984.

Kitawaki N., Honda M., Itoh, K, (1984), "Speech-quality assessment methods for speech-coding systems," *IEEE Communications Magazine,* vol.22, no.10, pp.26-33, October 1984.

Milner, B., Semnani S. (2000), "Robust speech recognition over IP networks," *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on* , vol.3, no., pp.1791-1794 vol.3, 2000.

Perkins M.E., Evans K. Pascal D., Thorpe L.A (1997), "Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm-ITU-T G.729," *IEEE Communications Magazine,* vol.35, no.9, pp.74-81, Sep 1997.

Table 1. Mean Opinion Score

MOS [%]	MOS	ITU Quality Scale
81 – 100	5	Best
61 – 80	4	High
41 – 60	3	Medium
21 – 40	2	Low
0 – 20	1	Poor

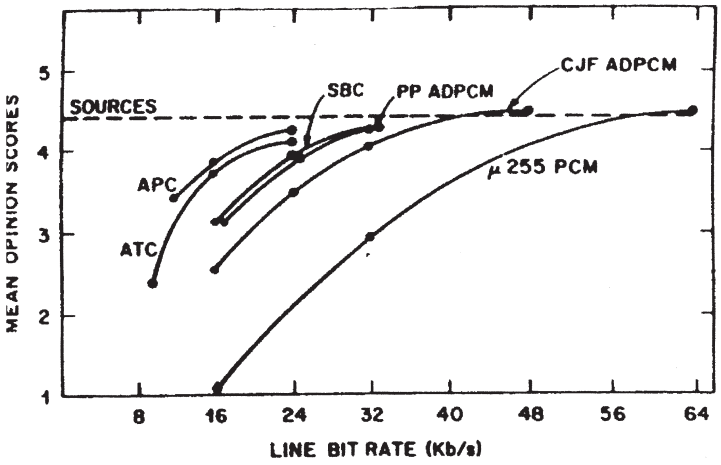

Figure 1. MOS values for different coding methods (Atayero, 2000)

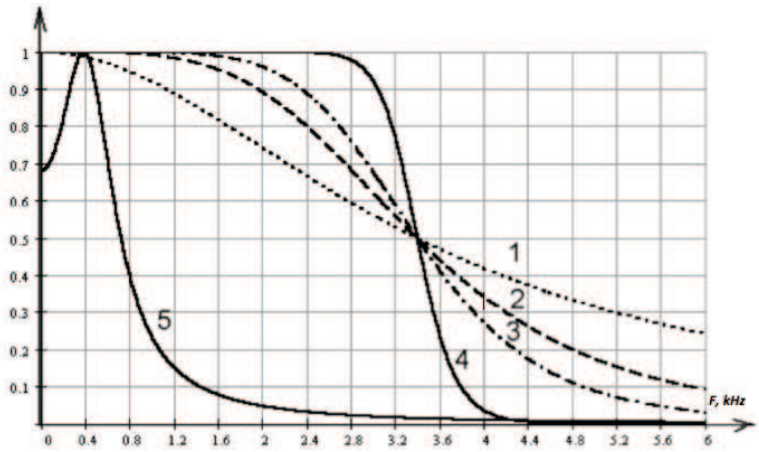

Figure 2. Filter transfer functions and normalized spectrum $for\ m = 1, 2, 3, 10\ and\ \alpha = 1.0\ kHz$.

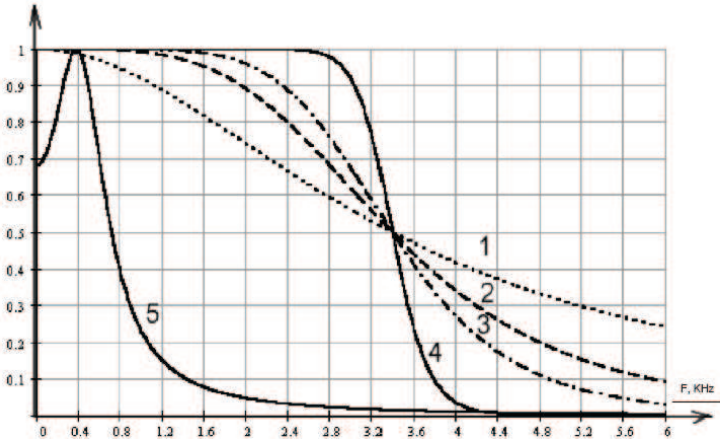

Figure 3. Filter transfer functions and normalized spectrum $for\ m = 1, 2, 3, 10\ and\ \alpha = 1.5\ kHz$.

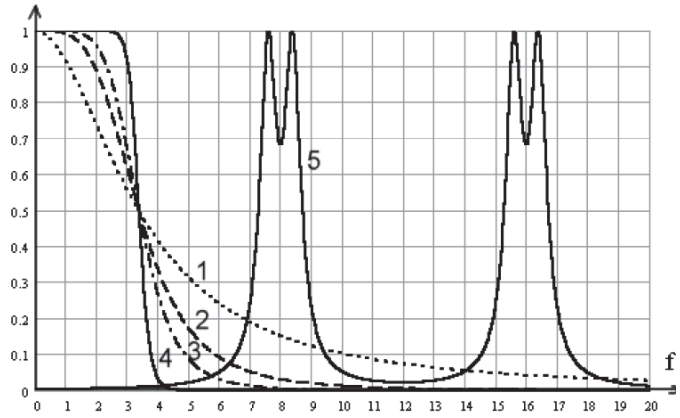

Figure 4. Filter transfer functions and normalized spectrum $for\ m = 1, 2, 3, 10\ and\ \alpha = 2.0\ kHz$.

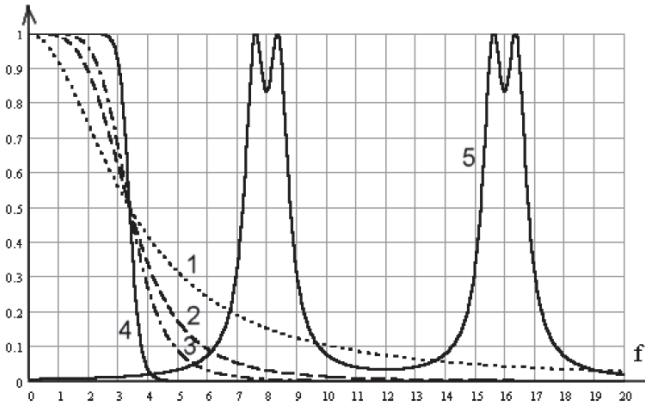

Figure 5. Filter transfer functions and normalized spectrum $for\ m = 1, 2, 3, 10\ and\ \alpha = 2.5\ kHz$.

For Figures 2 through 5: 1–filter transfer function for $m = 1$; 2–filter transfer function for $m = 2$; 3–filter transfer function for $m = 3$; 4–filter transfer function for $m = 10$; 5–normalized spectrum $S_1(\omega)$.