INFORMATICA, 2009, Vol. 20, No. 4, 487–498

An Investigation of the Perceptual Value of Voice Frames

Algimantas KAJACKAS, Aurimas ANSKAITIS
Telecommunications Engineering Department, Vilnius Gediminas Technical University
Naugarduko 41, LT-44404 Vilnius, Lithuania
e-mail: algimantas.kajackas@el.vgtu.lt; aurimas.anskaitis@el.vgtu.lt

Received: October 2008; accepted: December 2008

Abstract. It is well known, the voice segments and coincident data packets are not equally valued and significant for decoding and comprehension of speech signal. Some lost segments may only slightly worsen audible quality, while others cause strong distortion of the speech signals. Despite this, the feature of different importance of different voice segments in current generation of digital voice transmission systems is not fully used. There is a fundamental problem with discrimination of different importance and value of voice frames. In this paper the concept "of value of voice frame" is introduced, the metric and means for evaluation and measurement of voice frame value are proposed and also results of the measurements of voice frames value are presented.

Keywords: value of voice frame, PESQ measure.

1. Introduction

The significant growth in the volume of multimedia service and data transmission do not reduce the importance of voice and speech communications. New areas of research emerged when modern speech processing systems and telecommunications systems where human communicates with a computer were integrated. Well known are voice dialogues for booking tickets, catalog ordering, for generating text responses to e-mail received via a voice interface, etc.

The human speech is an analogue physical signal $v(t)$ that varies slowly in time. Current communication and voice processing systems means digitize voice signal, divide it into segments or frames of length T_f, and encode these segments. The ith segment or frame of voice signal $v(t)$ is depicted as code V_i and transmitted as ith data packet C_i

$$v(t) \Rightarrow V_i, \quad \text{where } t_0 + (i-1)T_f \leqslant t \leqslant t_0 + iT_f. \tag{1}$$

A wide variety of speech coders are available currently (Chu, 2003). In each one a distinct speech processing algorithm incorporated, different properties of the underlying voice are used. It is important that the duration of speech coding frame T_f, the code V_i of ith frame, and sample of transmitted information depend on the particular coder used.

By transmission of voice packets over mobile link or over the internet, some packets may be distorted severely or the delay of transmission may be too long to accept them.

In both cases such packets are discarded or dropped. With TCP for example, dropped packets cause a delay, but not loss. By typical transmission of voice dropped packets cause voice signal distortion and loss of some quantities of data and information.

For this work it is important that the voice segments and coincident data packets are not equally valued and significant for decoding and comprehension of speech signal. A lot of time segments of voice are perceptually irrelevant. It is well known, that some lost segments may only slightly worsen audible quality, while others cause strong distortion of the speech signals. Despite this, the feature of different importance of different voice segments in current generation of digital voice transmission systems is not fully used. The simplest concept of segmental importance is used in voice activity detectors for discriminating waveform segments as being one of "speech" or "silence". The concept of different value of voice components partially used in GSM and AMR voice coders (3GPP TS 26.090). But by transmission of voice all voice data packets are equally protected by error correcting codes. Consequently, the packet losses occurs independent from voice frame value or importance.

It will be observed that there are several recently published works that proposed to protect the important segments of speech by some special error correcting codes (Paraestholm *et al*., 2004) or by priority-marking (De Martin, 2001; Qiao *et al.*, 2004). Even though there are fundamental problem with discrimination of different importance and value of voice frames.

In this paper it is tried to analyze value of separated voice frames. The main purpose of this work is the definition of the value of voice frames, the development of the metrics and means for evaluation of voice frame value and also the evaluation of value (importance) of voice frames.

2. Value Analysis Fundamentals and Metrics

The concept of value is a "broad category". It was investigated by philosophers, psychologists, economists, etc. From value theory the concepts of intrinsic value and instrumental value came. The authors of this work are not brave enough to touch the concept of intrinsic value of voice, that a voice has "in itself". We choose the instrumental value which is the value of objects, as means of achieving something else[1].

There are a lot of published papers concerning segmental structure of speech, of words and phrases as well importance of phonological units of the language, such as phonemes, vowels, consonants. For example, consonants are known to be important in the intelligibility of speech even though they represent a relatively small fraction of the signal energy (Kazlauskas, 1999).

Current voice coders digitizing analog *voice* signal divide it into segments independent of phonological structure. The length T of voice frame typical is 10–30 ms and is shorter of phonological units of the language. Thus it is not possible to apply knowledge of phonological language structure while analyzing frame value of coded sound.

[1] `http://en.wikipedia.org/wiki/Instrumental_value`.

The key problem for applications is the definition and evaluation metrics of importance or value of some voice segment.

There are also many points of view concerning issues of value and importance of voice segment. The simplest and natural criterion for discrimination between speech and silence is energetic (Kazlauskas, 1999). But there are well known many voice activity detectors which use more sophisticated criteria (Prasad *et al.*, 2006). Another concept is used to classify a frame as being one of voiced or unvoiced (Kulesza *et al*., 2006). An advanced wide variety concepts of value for voice segment is characterized as the perceptual or communicative importance with is determined via an understanding of human physiology and cognitive psychology. Various speech recognition metrics (Lipeika and Lipeikienė, 2008) may be employed for the classification of voice segment importance.

The concept of perceptual importance has been used in various contexts too. Some examples: perceptual importance of voiced/unvoiced frames, perceptual importance of a pitch, the perceptual importance of formant and formant frequencies. These are examples when importance is estimated in a subjective way. For use in technical applications such as objective assessment of speech coders and transmission systems, common metrics are speech intelligibility and perceived voice quality. Let's say we have some criteria Q for assessing speech intelligibility or voice quality, which is defined as a functional $Q = \Phi[v(t)]$ of the voice signal $v(t)$.

The value of a particular voice frame in this paper is expressed in terms of the distortion that would be introduced by its loss. Strictly speaking, the value of ith frame is evaluated by estimating the impact on the criterion Q when this frame is lost

$$V_i = \alpha \cdot \Delta Q = \alpha \cdot \big[Q_0 - Q(i)\big], \tag{2}$$

where Q_0 is the initial quality of speech waveform some segment, $Q(i)$ is the quality value obtained when ith frame was lost, and α is the coefficient of proportionality.

The quality of transmitted speech in telecommunications systems depends on many technical factors such as channel bandwidth, signal level, echo, delay, signal-to-noise ratio, and codec type. The simplest criterion of quality may be signal-to-noise rate.

The most natural way of estimating the perceived voice quality Q is through use of subjective testing. The International Telecommunications Union (ITU) has developed a number of recommendations for subjective testing including the ITU-T recommendation P.800 which defines the Mean Opinion Score (MOS) as one important metric for subjective determination of transmission quality. However it is obvious that listening tests are subjective and extremely time-consuming. There are few objective methods for voice quality evaluation. Some are described as ITU-T recommendations P.561, P.563, P.861. Currently a reliable and objective voice quality measurement alternative is considered the so-called Perceptual Evaluation of Speech Quality (PESQ) method (ITU-T Rec. P. 862, 2001).

PESQ algorithm as the ITU-T speech quality standard P.862 is created according to the model of human listening. The PESQ is an intrusive measurement algorithm. Quality prediction is based on a comparison of a reference and a degraded signal (Fig. 1).

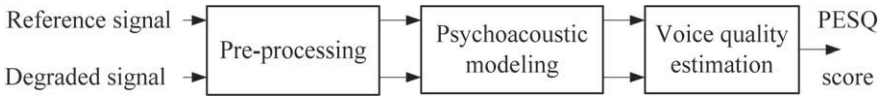

Fig. 1. Structure of PESQ algorithm.

The structure of PESQ can be described as follows. After preprocessing, such as level- and time-alignment, both the reference and degraded speech signal are transformed into a psychoacoustic representation which models the properties of the human auditory system. In the perceptual domain, the signals are compared and a speech quality estimate is calculated which corresponds to a subjective mean opinion score (MOS) ranging from 1 (bad) to 4.5 (excellent) in most cases. Some more detailed information on the psychoacoustic model of PESQ is described in paper of Beerends *et al*. (2002).

At first the degraded signal is adjusted to the original signal. Next, the perceptual difference between the original signal and the degraded version is calculated. Finally, PESQ calculates the perceived speech quality of the degraded signal. The correlation between MOS and PESQ scores over a broad range of speech data is 0.935 (ITU-T Rec. P. 862, 2001).

In this paper PESQ was chosen for analysis of importance and value of voice segments because it is well analyzed and respected objective method for voice quality evaluation. It is important to note that there are similar research using standard PESQ algorithm (Hoene *et al.*, 2003). We began our experiments in the same way. We note that PESQ measurements results when using short speech segments ($\sim$1 s) are unstable. Formally this can be explained very easy. The recommended usage conditions of PESQ require signals to be 8–30 s long (ITU-T Rec. P. 862.3, 2005).

3. Brief Presentation and Analyze of PESQ

The difference between the degraded $v^d(t)$ and the reference signals $v(t)$ is processed through several steps using special time segmentation as shown in Fig. 2. The basic unit is a "phoneme" with size of $T_{\rm ph} = 32$ ms. Overlap between successive phoneme windows is 50 percent. 20 overlapped phonemes amount to one $T_{\rm sl} = 320$ ms long syllable.

The first algorithm step carries integration over frequency to estimate for every phoneme the measures of the perceived disturbance. Symmetric disturbances $D_{\rm sn}$ and

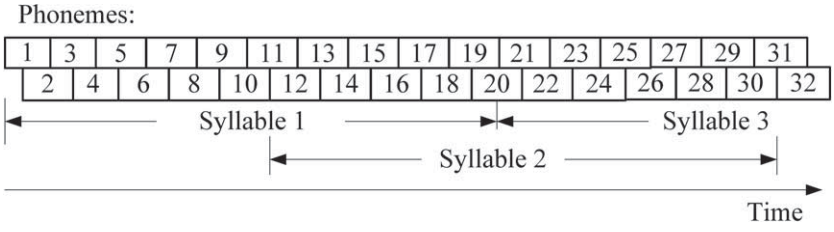

Fig. 2. PESQ time segmentation.

asymmetric disturbances D_{an} are calculated for every phoneme in each syllable. At the next step, signal is transformed into a psychoacoustic representation – phoneme disturbance values D_{sn} and D_{an} are aggregated for every syllable using the following formulas:

$$L_{\mathrm{DS}} = \left(\frac{1}{20} \sum_{n=1}^{20} D_{\mathrm{sn}}^{6} \right)^{1/6}, \qquad L_{\mathrm{DA}} = \left(\frac{1}{20} \sum_{n=1}^{20} D_{\mathrm{an}}^{6} \right)^{1/6}. \tag{3}$$

Finally, the algorithm carries out integration over the entire speech signal measurement window – the aggregation of syllable disturbances using typical mean square algorithm:

$$d_{\mathrm{sym}} = \left(\frac{1}{N} \sum_{i=1}^{N} L_{\mathrm{DS}}^{2}(i) \right)^{1/2}, \qquad d_{\mathrm{asym}} = \left(\frac{1}{N} \sum_{i=1}^{N} L_{\mathrm{DA}}^{2}(i) \right)^{1/2}. \tag{4}$$

Here N is number of syllables in PESQ measurement window T. The final PESQ score is calculated as follows:

$$Q_{\mathrm{PESQ}} = 4.5 - 0.1 \cdot d_{\mathrm{sym}} - 0.0309 \cdot d_{\mathrm{asym}}. \tag{5}$$

Value of the ith voice frame when using definition (2) and PESQ measure (5) can be calculated by following formula:

$$V_{\mathrm{PESQ}} = 4.5 - Q_{\mathrm{PESQ}} = 0.1 \cdot d_{\mathrm{sym}} - 0.0309 \cdot d_{\mathrm{asym}}. \tag{6}$$

The aggregation of syllable disturbances using Eqs. (3) and (4) is well-justified when main factors affecting degradation of speech quality under typical conditions of application are stationary background noise or coding distortions. Disturbances D_{sn} and D_{an} arising from these factors have almost equal value over the entire duration of syllables.

In modern communication systems voice is transmitted in frames with duration of 10 or 20 ms. It is easy to see from Fig. 2 one such lost frame may affect two or three PESQ phonemes. It depends on the position of the lost frame in a PESQ measurement window. Affected phonemes may impact two or three syllables in their turn. It also depends on the position of the lost frame.

Calculation of PESQ score under such circumstances may be characterized as follows:

1. Phoneme disturbance values D_{sn} and D_{an} will definitely be not zeros only for phonemes which falls into the same interval as the frame under investigation. All other phonemes will have zero disturbances in (3). Syllables containing non-zero disturbance phonemes identified in the above clause will also have non-zero disturbance values,

2. Moving of PESQ measurement window's starting time will re-distribute the impact of lost frame between phonemes and syllables. Re-distribution of impact of lost frame between phonemes and syllables creates dependence of measurement result on position of the lost frame in the measurement window (Kajackas *et al.*, 2008),

3. The aggregation of syllable disturbances is done using typical mean square algorithm (4) where one lost frame can influence from 1 to 3 syllables. Other components

of (4), i.e., $L_{\mathrm{DS}}(i)$ and $L_{\mathrm{DA}}(i)$, where $i > 3$, will be equal to zero. In this case Eqs. (4) can be transformed into:

$$d_{\mathrm{sym}} \approx \frac{L_{\mathrm{sym}}}{\sqrt{N}}, \qquad d_{\mathrm{asym}} \approx \frac{L_{\mathrm{asym}}}{\sqrt{N}}, \tag{7}$$

where N is a number of syllables in measurement window. From here follows if measurement window is longer than tree syllables (720 ms) then quality degradation score is artificially decreased.

So we have obvious collision:

1. When duration of measurement window is small (1–3 syllables) we get uncertain result because of position of lost frame in the measurement window.
2. When duration of measurement window is longer (>3 syllables) degradation score is artificially decreased.

This collision in our work is solved using special synthesized signal.

4. Signal Synthesis for the Evaluation Value of Voice Frame

To investigate the value of voice frame we choose a signal $v_t(t)$ of relatively short duration T_w. In order to reduce the uncertainty of PESQ measurements, we propose using specially composed test signal. Its design is based on the provision of ITU-T Rec. P.862 that "real speech test signal may be constructed by concatenating short fragments of real speech while retaining a representative structure of speech and silence". Using this provision, an extended signal consisting of periodically J times replicated original voice segments $v_t(t)$ may be constructed:

$$v_\Sigma(t) = \begin{cases} v_t(t - jT_r), & \text{if } (j-1)\cdot T_r \leqslant t \leqslant j \cdot T_r, \\ 0, & \text{if } t \geqslant JT_r \text{ or } t < 0, \end{cases} \tag{8}$$

where T_r is a replication period, $T_r \leqslant t \leqslant JT_r$, $j = 1, 2, \ldots, J$. Formula (8) describes the reference signal $v_\Sigma(t)$. To investigate value of ith frame within the measurement window T_w, the "degraded" signal is constructed in two stages.

In the first stage signal of duration T_w is created by deletion of ith frame according to rule described follows:

$$v_t^d(t) = \begin{cases} v_t(t), & \text{when } t_0 \leqslant t \leqslant t_0 + (i-1)\cdot T_f, \\ 0, & \text{when } t_0 + (i-1)\cdot T_f \leqslant t \leqslant t_0 + i \cdot T_f, \\ v_t(t), & \text{when } t_0 + i \cdot T_f \leqslant t \leqslant t_0 + I \cdot T. \end{cases} \tag{9}$$

In the second stage extended degraded signal $v_\Sigma^d(t)$ is made according to formula (8) by replacing the original signal $v_t(t)$ with degraded one $v_t^d(t)$. In reality when forming signals according to (8) and (9) formulas, operations are performed not on signals themselves but on their codes V_i as follows from Eq. (1).

The replication period in (8) shall be chosen as $T_r \geqslant T_w$. This way it is possible to write:

$$T_r = T_w + \Delta T, \tag{10}$$

where ΔT is the additional increase of replication period. In this manner we achieve randomization of position of lost frame within extended measurement window.

The choice of the initial voice signal and measurement window T_w length is arbitrary. The following logic was applied while choosing this parameter. Subjective evaluation of voice distortion is possible when signal contains a word or few words. Thus the shortest duration T_w can be chosen similar to the duration of a word in the investigated language. The articulatory durations for words and syllables are estimated in Mueller *et al*. (2003), Norkevičius and Raškinis (2008). The average duration of isolated words is 499 ms (stdev. is 104 ms) and the average duration for words in memorized sequences is 273 ms. (stdev. is 69 ms). It can be seen that duration of words may be different. For our experiments we used 640 ms.

5. Experimental Test Bed and its Verification

For analysis of value of a single voice frame in some sentence an experimental test-bed was developed, as shown in Fig. 3. The PESQ measurer compares two extended signals: extended reference signal $v_\Sigma(t))$ and extended degraded signal $v_\Sigma^d(t)$. Both signals constructed according to formulas (8) and (9), where the signal $v_t(t)$ is exchangeable by performing further experiments.

Voice segment used in analysis $v_t(t)$ is encoded in E. Encoded signal travels two ways. At the first one signal is decoded at decoder D1 and extended in recirculator R1.

The second way modified replica $v_t^d(t)$ is obtained by submitting the encoded signal $v_t(t)$ through the control block which imitates a loss of ith frame. Further down the way the degraded signal $v_t^d(t)$ is obtained by decoding bits from control block. Recirculator R2 finally generates signal $v_\Sigma^d(t)$ used for quality calculation. In our simulations the number of iterations in both recirculators $J = 8$ was used. This way PESQ measurement precision is improved more than 10 times and calculation overhead is acceptable for real time applications.

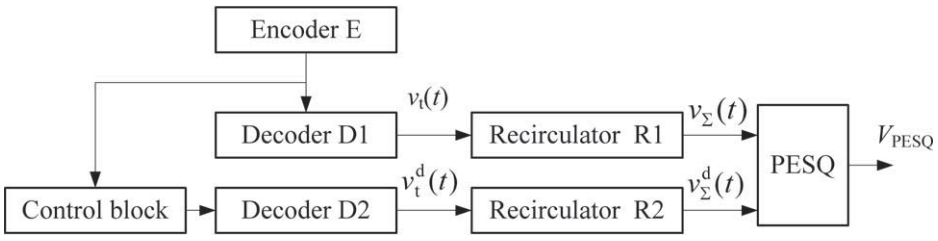

Fig. 3. Experimental test-bed.

The experiments were performed using AMR-12.2 (3GPP TS 26.090) as encoder E and decoders D1 and D2. The encoding and decoding of the original and degraded signals using codec of the same type eliminates the influence of the codec on voice quality and value of selected frame.

The significant peculiarities of AMR coder could be listed as follows:

- The codec divides input voice signal into 20 ms frames.
- The error concealment algorithm (3GPP TS 26.091) includes active and passive concealment in the form of parameter repetition/substitution, and attenuation/muting. This approach sometimes can be damaging to speech quality and intelligibility, and can influence variations of the frame value. Therefore the measurement of voice frame value is conditional, dependent on the choice of voice codec.

6. Verification of Test Bed

The effectiveness of the proposed signal synthesis and test bed was confirmed by many subsequent experiments performed with different voice signals.

The first experiment was performed to verify the uncertainty of measurements. The absolute uncertainty of frame value V_{PESQ} measurements in measurement window $T_{w.} = 640$ ms using original (not extended) signals $v_t(t)$ and $v_t^d(t)$ varied from 0.07 to 1.3. The same experiment performed using extended signals $v_\Sigma(t)$ and $v_\Sigma^d(t)$ gave variation of uncertainty only between 0.001 and 0.09.

The second experiment was performed to verify the impact of measurement window length on measurement results. This experiment was performed using two windows: 0.64 and 1.28 s. The second signal is longer but the first parts of signals chosen identical. The frame losses are imitated in the first parts of signal. The measurement results $V_{\mathrm{PESQ}}(1.28)$ and $V_{\mathrm{PESQ}}(0.64)/\sqrt{2}$ are shown at Fig. 4. (The value of $V_{\mathrm{PESQ}}(0.64)$ is multiplied by $1/\sqrt{2}$ because of (7).)

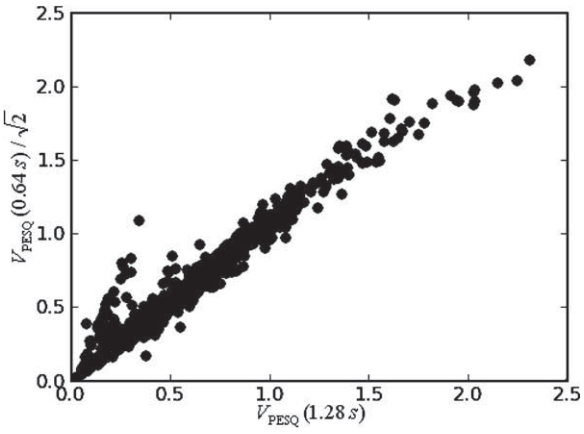

Fig. 4. The influence of measurement window length on value of frame.

The correlation coefficient of this measurement is calculated 0.92. So if we want to compare measurements of different durations such scaling of data is mandatory.

7. Distributions of Values of Voice Frames

The first experiment performed to visually show the variation of value of frames in an isolated word. The voice signal by pronouncing of lithuanian word "*lova*" (bed) is shown in upper part of Fig. 5, corresponding frame values are shown in lower part of Fig. 5. The top value frames are those beginning speech burst. The frames in a long vowel have less value because frame substitution algorithm compensates well for such losses. Similar value distributions are observed with other words too.

The next experiment will try to determine the statistical value of a frame in English language. The experiment simulates frame loss in a long signal and calculated V_{PESQ} score for the frame using $T_w = 640$ ms window. In the same way frame value was measured for all 3000 frames. The distribution of frame values and some statistics were calculated from the measured data.

Fig. 6 shows the empirical distributions of frame value for two voice records. The first record was CNN news and it contains noticeably very fast speech (solid line). The second record is interview about sport and it may be characterized as "slow" (dotted line). Statistical parameters of values for "fast" and "slow" speech records are given in Table 1.

It can be concluded that distributions of frame values for various voice records are significantly different. The distributions are depending on nature of speech. The mean frame value for "fast" speech is 0.94 and mean for "slow" speech is 0.55. The percent of

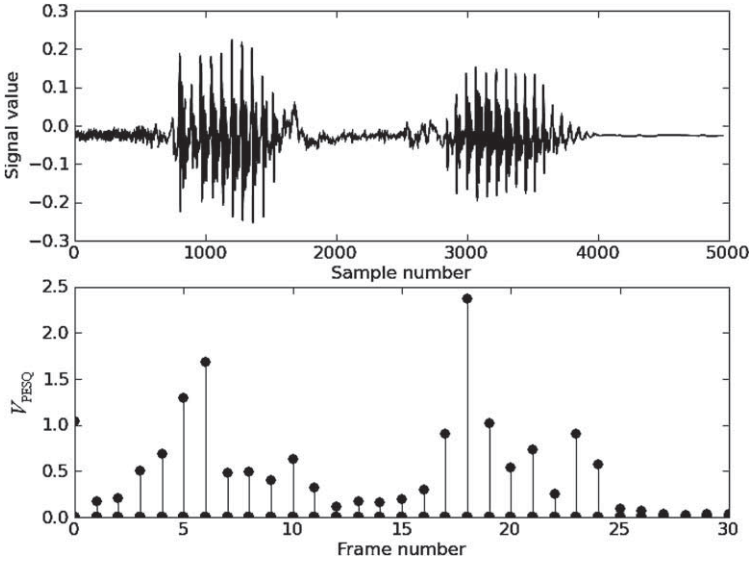

Fig. 5. Example of voice frame value in Lithuanian word "sofa" (sofa).

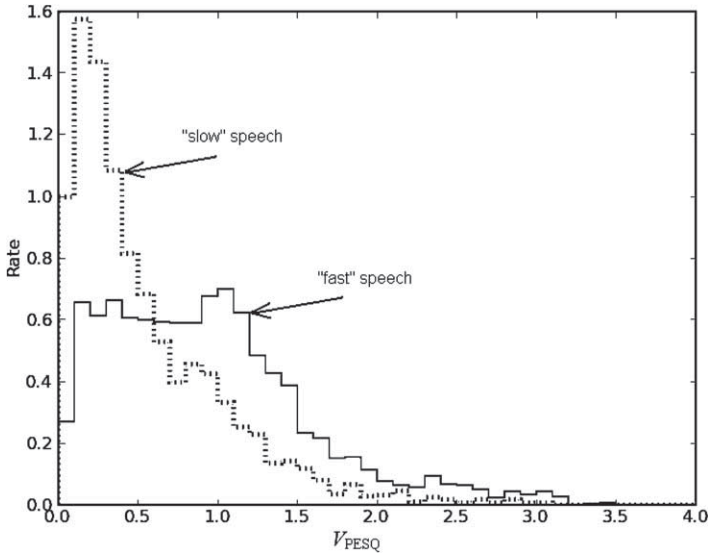

Fig. 6. Empirical distributions of frame value for “fast” and “slow” speech records.

Table 1
Statistical parameters for different speech types

Parameter	Speech type	
	“Fast”	“Slow”
Mean	0.94	0.551
Standard deviation	0.627	0.487

low valued frames is higher for “slow” speech. The percent of very high valued frames (higher as 3.0) is not big for both voice records.

The proposed metrics of voice value create the possibility to classify frames to the classes of different value. Table 2 shows an example of division of frame values into seven classes. In this table there are empirical probabilities for particular segments of frame value presented. These empirical probabilities are denoted using $P(r_1, r_2)$ notation, where r_1 and r_2 are range of frame value limits.

This way differentiated protection of frames may be implemented. Distinctly labeled frames shall be coded by different error protecting codes.

8. Conclusions

This work is devoted to the study on perceptual importance of voice frames. Recognizing the unequal perceptual importance of voice frames, we propose the measure of frame value based on Perceptual Evaluation of Speech Quality (PESQ). Analysis of the structure

Table 2
Empirical probabilities for frame value to fall into ranges

Probability	Speech type	
	"Fast"	"Slow"
$P(0, 0.5)$	0.279	0.570
$P(0.5, 1.0)$	0.304	0.240
$P(1.0, 1.5)$	0.262	0.105
$P(1.5, 2.0)$	0.088	0.031
$P(2.0, 2.5)$	0.036	0.012
$P(2.5, 3.0)$	0.022	0.005
$P(3.0, +\text{inf})$	0.008	0.035

of PESQ yielded a method of its application to short signals with overall duration under 1 s. This method avoids some pitfalls of original PESQ algorithm. The proposed measure of value is conditional on the choice of voice codec.

In this paper examples of evaluation of value of voice frames are presented. The empirical probability distribution of frame values depends on the type of speech analyzed. It can be seen that part of significant or high evaluated frames is not high. But by loss of these frames the experienced degradation of voice quality will be high. Consequently these high evaluated frames shall be better protected with stronger channel codes than the less evaluated.

The proposed metrics of voice value create the possibility to classify frames to the classes of different value. This way differentiated protection of frames may be implemented. The proposed metrics of voice value also allows evaluate voice quality of transmitted voice objectively.

Acknowledgment. We would like to thank Lithuanian State Science and Studies Foundation for partial support for this work.

References

Beerends, J.G., Hekstra, A.P., Rix, A.W., Hollier, M.P. (2002). Perceptual evaluation of speech quality (PESQ): the new ITU standard for end-to-end speech quality assessment, Part II: Psychoacoustic model. *Journal Audio Eng. Soc.*, 50(10), 765–778.

Chu, W.C. (2003). *Speech Coding Algorithms*: *Foundation and Evolution of Standardized Coders*. Wiley.

De Martin, J.C. (2001). Source-driven packet marking for speech transmission over differentiated services Networks. In: *Proceedings of IEEE ICASSP*, pp. 753–756.

ITU-T Rec. P.800. (1996). Methods for subjective determination of transmission quality.

ITU-T Rec. P. 862 (2001). Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

ITU-T Rec. P. 862.3 (2005). Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P. 862.2.

Hoene, C., Rathke, B., Wolisz, A. (2003). On the importance of a VoIP packet. In: *Proc. of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems*. Germany, pp. 55–62.

Kajackas, A., Anskaitis, A., Guršnys, D. (2008). Peculiarities of testing the impact of packet loss on voice quality. *Electronics and Electrical Engineering*, 2(82), 35–40.

Kazlauskas, K. (1999). Noisy speech intelligibility enhancement. *Informatica*, 10(2), 171–188.
Kulesza, M., Szwoch, G., Czyżewski, A. (2006). High quality speech coding using combined parametric and perceptual modules. In: *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 13, pp. 244–249.
Lipeika, A., Lipeikienė, J. (2008). On the use of the formant features in the dynamic time warping based recognition of isolated words. *Informatica*, 19(2), 213–226.
Mueller, S.T., Seymour, T.L., Kieras, D.E., Meyer, D.E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, 29(6), 1353–1380.
Norkevičius, G., Raškinis, G. (2008). Modeling phone duration of Lithuanian by classification and regression trees, using very large speech corpus. *Informatica*, 19(2), 271–284.
Paraestholm, S., Jensen, S.S., Andersen, S.V., Murthi, M.N. (2004). On packet loss concealment artifacts and their implications for packet labeling in voice over IP. In: *IEEE International Conference on Multimedia and Expo* (*ICME*), pp. 1667–1670.
Prasad, R.V., Vijay, S., Shankar, H.N., Pawelczak, P., Niemegeers, I. (2006). Voice activity detection for VoIP – an information theoretic approach. In: *GLOBECOM. IEEE*, pp. 1–6.
Qiao, Z., Sun, L., Heilemann, N., Ifeachor, E. (2004). A new method for VoIP quality of service control use combined adaptive sender rate and priority marking. In: *IEEE Communications Society*, pp. 1473–1477.
3GPP TS 26.091, Universal Mobile Telecommunications System (UMTS). Mandatory Speech Codec speech processing functions; AMR Speech Codec; Error concealment of lost frames, ETSI.
3GPP TS 26.090, Universal Mobile Telecommunications System (UMTS); Mandatory Speech Codec speech processing functions; AMR Speech Codec; Transcoding functions, ETSI.

A. Kajackas received the degrees of candidate of science (PhD) and doctor of science in telecommunications (Dr. Hab.) from Sankt-Petersburg Institute of Electrical Engineering & Telecommunications in 1967 and 1974, respectively. From 1967 he has worked in Telecommunications Engineering Department at the Kaunas University of Technology, Lithuania. He is currently head of Department of Telecommunications Engineering at the Vilnius Gediminas Technical University. Ongoing research projects include wireless networks and quality of service in mobile networks.

A. Anskaitis received masters' degree from Vilnius Gediminas Technical University in 2005. From 2005 he works and prepares his doctoral thesis in Telecommunications Engineering Department at Vilnius Gediminas Technical University. Ongoing research projects include analysis of speech quality evaluation algorithms.

Balso segmentų suvokiamosios vertės tyrimas

Algimantas KAJACKAS, Aurimas ANSKAITIS

Yra gerai žinoma, kad balso segmentai ir juos atitinkantys paketai turi nevienodą vertę suprantamumui. Kai kurie prarasti paketai gali tik truputį sumažinti girdimą audio kokybę, kai kiti sukelia didelius pokyčius signale. Nepaisant šito, šiuolaikinėse balso kodavimo ir perdavimo sistemose informacija apie skirtingą paketų reikšmingumą nėra naudojama arba išnaudojama nepilnai. Balso signalų diskriminavimas pagal jų reikšmingumą yra fundamentali problema.

Šiame darbe įvesta sąvoka „balso paketo vertė“, apibrėžtos metrikos ir priemonės, kurios leidžia matuoti šią balso vertę.