VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Mohamad ITANI

ANALYSIS AND IMPROVEMENT OF MULTILINGUAL SPEECH CODING

DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
ELECTRICAL AND ELECTRONIC ENGINEERING (01T)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2011

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2006–2011.

Scientific Supervisor

Assoc Prof Dr Šarūnas PAULIKAS (Vilnius Gediminas Technical University, Technological Sciences, Electrical and Electronic Engineering – 01T).

VGTU leidyklos TECHNIKA 1858-M mokslo literatūros knyga
http://leidykla.vgtu.lt

ISBN 978-9955-28-815-2

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Mohamad ITANI

ĮVAIRIŲ KALBŲ KALBOS SIGNALŲ KODAVIMO ANALIZĖ IR GERINIMAS

DAKTARO DISERTACIJA

TECHNOLOGIJOS MOKSLAI,
ELEKTROS IR ELEKTRONIKOS INŽINERIJA (01T)

VGTU
LEIDYKLA
TECHNIKA
Vilnius 2011

Disertacija rengta 2006–2011 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas

doc. dr. Šarūnas PAULIKAS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, elektros ir elektronikos inžinerija – 01T).

Abstract

The thesis investigates the performance of voice codec's that uses linear predictive coding (LPC), over different languages. The English language has had a dominating influence on the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust. It is noted that the quality of a speech produced by voice codecs mainly is assessed using samples of English language. Investigations show that most low-rate speech coders show bias towards non-accented English. When the coders are used for heavily accented English or other languages, significant performance degradation is noted.

The main goal of the thesis is to investigates and propose ways for improvement of the performance of CELP voice codecs for coding multilingual speech. In order to achieve the goal, the following problems had been solved: performed analytical review on aspects of multilingual CELP voice coding; created speech records of Lithuanian language database according to ITU-T Rec. P.50 App. 1. that is suitable for tests of performance voice codecs; investigated performance of AMR and Speex codecs for coding non-English speech and proposed and proved ways for improvement of multilingual speech coding.

Chapter 1 reviews most of the work done in the field of influence of language on coding, by introducing CELP based codes, and previous researches about the topic in hand. As well, it concludes in formulating of main objective and tasks of present investigation.

Chapter 2 introduces the quality measuring techniques categorized by subjective and objective. Moreover, it contains our primary experiment on the codecs and quality assessment techniques.

Chapter 3 presents the construction of the Lithuanian speech database. After concluding the results in chapter 2, we found ourselves in need of a well formed speech database for Lithuanian language to carry out our experiments on.

Chapter 4 presents the experimental study on the performance of CELP based codes under different languages.

Chapter 5 suggests a solution the codecs to perform better under different languages.

Three scientific articles have been published in reviewed scientific publications by the author within scope of research: two publications in reviewed ISI Web of Science journals and one in other reviewed scientific editions. Also results of performed investigations were presented in 4 scientific conferences.

Reziumė

Disertacijoje tiriamas balso kodekų taikančių tiesinės prognozės mechanizmą našumas koduojant skirtingų kalbų kalbos signalus. Anglų kalba nuo seno dominavo ir darė didžiausę įtaką telekomunikacijų technikos raidai. Tai lėmė faktai, kad dauguma atradimų ir patobulinimų buvo padaryta angliškai kalbančiose šalyse. Taigi, yra labai tikėtina, kad šie patobulinimai yra lingvistiškai nepagrįsti. Yra pastebėta, kad balso kodekais atkurto kalbos signalo kokybė yra dažniausiai vertinama anglų kalbai. Tyrimai rodo, kad mažos duomenų spartos kodavime kitų, ne anglų kabos, kalbos signalai yra labiau sugadinami.

Pagrindinis disertacijos tikslas yra ištirti CELP balso kodekų našumą koduojant skirtingų kalbų signalus ir pasiūlyti būdus kodavimo pagerinimui. Siekiant šio tikslo reikia išspręsti šiuos uždavinius: atlikti analitinę CELP balso kodavimo koduojant skirtingų kalbų signalus analizę; sukurti lietuvių kalbos įrašų duomenų bazę pagal to ITU-T P.50 rekomendacijos 1 priedo analogą, kuri būtų tinkama balso kodekų testavimui; ištirti AMR ir Speex našumą dekoduoto balso kokybės prasme koduojant ne anglų kalbos signalus.

Pirmame disertacijos skyriuje yra atliekama kalbos įtakos balso kodavimo teorijoje apžvalga, apžvelgiamas balso kodekų veikimo principas, formuluojami disertacijos uždaviniai.

Antrame skyriuje pateikiamos kalbos signalo kokybės samprata, jos vertinimo būdai. Atliekami pirminiai kodekų kodavimo gerumo įvairioms kalboms tyrimai.

Trečias skyrius skirtas sukurtos lietuvių kalbos įrašų duomenų bazės aprašymui ir jos tinkamumo kodekų testavimui vertinimui.

Ketvirtame skyriuje aprašomi AMR ir Speex kodekų kokybės tyrimai koduojant anglų, lietuvių ir arabų kalbų signalus.

Penktame skyriuje pateikiamas kodekų kodavimo pritaikymo konkrečiai kalbai būdas – fiksuotos kodų lentelės adaptavimas konkrečiai kalbai. Parodoma, kad tai leidžia pagerinti Speex kodeko kokybę koduojant lietuvių ir arabų kalbų kalbos signalus.

Disertacijos tematika yra paskelbtos trys publikacijos recenzuojamuose mokslo leidiniuose: du straipsniai ISI Web of Science žurnaluose ir vienas tarptautinės konferencijos medžiagoje indeksuojamoje IEEE Xplore duomenų bazėje. Disertacijos rezultatai viešinti keturiuose mokslinėse konferencijose.

Notations

Symbols

$A(z)$	–	The Optimum Unquantized Inverse Filter (for subframe n);
$A_p(z)$	–	The Short Term Quantized Inverse Filter (order p);
$B(z)$	–	The Long Term Quantized Inverse Filter;
C	–	Codebook;
D	–	Spectral Distortions;
d_c	–	Euclidean Distance Between Codebooks;
ϵ	–	The Weighted Mean Squared Error;
G	–	Gain;
H	–	Lower Triangular Toeplitz Matrix;
$h(n)$	–	Impulse Response;
$S\left(e^{j2\pi k/L}\right)$	–	Original Speech Signal;
$\hat{S}\left(e^{j2\pi k/L}\right)$	–	Distorted Speech Signal;
$s[n]$	–	Original Signal;
$\hat{s}[n]$	–	Distorted Signal;
SNR	–	Signal to Noise Ratio;
σ	–	Dispersion;
p	–	Filter Order;

$\mathcal{T}^{(i)}$ – Training Set;
$\boldsymbol{v}$ – Excitation Vector;
x_i and y_i – Data.

Abbreviations

3SQM – Single Sided Speech Quality Metric;
ACELP – Algebraic Code Exited Linear Prediction;
AMR – Adaptive Multi Rate;
BFI – Bad Frame Indication;
CELP – Code Exited Linear Prediction;
CHD – Channel Decoder;
CHE – Channel Encoder;
CNG – Comfort Noise Generation;
DRT – Diagnostic Rhyme Test;
DTX – Discontinuous Transmission;
GSM – Global System for Mobile communications;
ITU-T – International Telecommunication Union, Telecommunication Sector;
LPC – Linear Prediction Coding;
LRN – Lithuanian Radio News Agency;
MOS – Mean Opinion Score;
MRT – Modified Rhyme Test;
PCM – Pulse Code Modulation;
PESQ – Perceptual Evaluation of Speech Quality;
PLMN – Public Land Mobile Network;
PSTN – Public Switched Telephone Network;
RX – Receive;
SCR – Source Controlled Rate;
SD – Spectral Distortions;
SNR – Signal to Noise Ratio;
SPD – Speech Decoder;
SPE – Speech Encoder;
TC – Transcoder;
TX – Transmit;
UDP – Unacknowledged Datagram Protocol;
VAD – Voice Activity Detector;
VoIP – Voice over Internet Protocol.

Contents

List of Figures

List of Tables

Introduction

The Investigated Problem

Technologies of telecommunication networks in these days migrate to multidimetional heterogeneity as they are widely used for multimedia applications. Until now communication by speech remains the most widely used for exchange of information among humans. The QoS of speech communication strongly depends on in electronic telecommunication system employed voice codec and its parameters (Kajackas and Anskaitis, 2009).

QoS in telecommunications networks and systems is analyzed by many authors in various aspects (Anskaitis, 2009; Cavander, *et al.* 2008; Joskowicz, *et al.* 2009; Kajackas, *et al.* 2009).

However, human speeches (sounds) differ from one language to another (Paulikas and Navakauskas, 2005). For example, with comparison to the English language, the Lithuanian language uses many vowels, while the Arabic language uses only three. The English language has had a dominating influence in the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust. Research carried out in paper, (Donald and Gersho, 1986) shows that coders interact with individual voices so that speech is degraded differentially for different talkers.

In modern electronic communication systems in the design of low-bit-rate speech coding algorithms, language variability is often considered to be of secondary importance. It is surprising to find that no comprehensive studies have ever been carried out on the role of language in speech coding. Parray, *et al.* (2000), found that quantization distortion is not uniform across languages and it influences codecs codebook performance and overall quality of decoded speech.

Today in electronic telecommunications sector a lot of companies that design voice communication products and regulatory institutions admit importance of language on performance of voice communication systems. For example, over 20 languages (mostly from European countries) that are listed in ITU-T Rec. P.50 Appendix I were extensively used in tests of AMR codec that were chosen by 3GPP as default voice codec in 3G mobile communications (3GPP TR 26.975 V5.0.0. 2002).

It is important to normalize the voice codecs performance over different non-English languages in modern electronic voice communication systems to give possibilities for users to obtain better QoS of voice communication services.

Topicality of the Research Work

Unfortunately, Arabic and Lithuanian language, as other languages of minor nationalities, has left aside of investigations of their influence on modern voice codecs performance. One of the reasons for not including these languages in codecs performance and other tests is lack of appropriate speech records databases. It should be noticed that in Lithuania there is a few Lithuanian speech records databases created by Institute of Mathematics and Informatics, Lithuanian Radio News (LRN0 and LRN0.1), Vytautas Magnus University (VDU–RTG). Also, Kaunas University of Technology together with University of Vilnius created LITGIS database (records of Lithuanian digits sequences). However, these speech records databases are intended to use for Lithuanian speech recognition but not for testing of voice communication systems.

So, there arises the need for creating a Lithuanian speech record database suitable for analysis of the performance of communication systems. Also it is important to investigate the voice codecs performance over different non-English languages in order to find out the influence of specific language on quality of coded speech.

Also it is important to investigate the voice codecs performance over different non-English languages in order to find out the influence of specific language on quality of coded speech and to propose of methods that provide techniques for improving the voice codecs performance for specific language. It

is essential to modern electronic voice communication systems that will give possibilities for users to obtain better QoS of voice communication services.

The Object of Research

The *object* of present research is analysis and ability to improve the performance of modern CELP voice codecs over different non-English languages.

The Aim of the Work

The aim of the work is to investigate and propose ways for improvement of the performance of CELP voice codecs for coding multilingual speech.

The Tasks of the Work

In order to achieve the goal, the following problems had to be solved:

1. Perform analytical review on aspects of multilingual CELP voice coding.
2. Create the speech records of Lithuanian language database according to ITU-T Rec. P.50 App. 1. suitable for tests of performance voice codecs.
3. Investigate the performance AMR and Speex codecs for coding Arabic and Lithuanian speech.
4. Propose and prove techniques for improvement codecs' quality in of multilingual speech coding.

Research Methodology

To investigate the *object*, the following *research methods* are chosen:

- Statistical analysis and analytical approach.
- Computer modeling and simulation (Matlab, C++).
- Active experimental tests.

Scientific Novelty and its Importance

The aspects of scientific novelty on theoretical and experimental investigation of the performance of CELP voice codecs over different languages are as follows:

1. Assessment of quality in sense of decoded speech obtained during experimental tests of AMR and Speex codecs' for decoding speech of English, Lithuanian and Arabic languages.
2. PESQ and 3SQM speech quality evaluation algorithms assessment obtained during experimental tests of their suitability to evaluate a quality of speech.
3. The Speex codec codebook adaptation technique that improves its performance in sense of speech quality while coding Lithuanian and Arabic speech.

Practical Significance of Achieved Results

Created speech records of Lithuanian language database according to ITU-T Rec. P.50 App. 1. is applicable for tests of performance voice communication systems.

The experimental test results of performance of CELP codecs may be used for specific voice codec performance evaluation and adaptation.

The adapted Speex codec's codebook could be employed for use in VoIP communication software.

The results of performed research were used in one scientific project supported by Lithuanian State Science and Study Fund, *Investigation of Lithuanian Language Influence on Coding of Speech signals*, Scientific project Nr. T–89/09, (Reg. Nr. T–09022).

The Defended Statements

The following statements based on the results of present investigation may serve as the official hypotheses to be defended:

1. The quality of Arabic and Lithuanian speech coding using CELP based voice codecs depends on specific language.
2. AMR and Speex codecs quality evaluated using objective and subjective measures show bias towards English language.

3. Fixed codebook adaptation for Lithuanian and Arabic language improves quality of speech coded by Speex codec and Speex codec's performance stability.

Approval of the Results

3 scientific articles have been published in reviewed scientific publications by the author within scope of research:

- 2 in reviewed ISI Web of Science journals;
- 1 in other reviewed scientific editions.

The author has made 4 presentations at scientific conferences:

- The 9th IEEE International Symposium on Signal Processing and Information Technology. December 14–17, 2009 – Ajman, UAE.
- International conference "ELECTRONICS" in 2007 and 2008.
- Conference "Science – Future of Lithuania" 2007.

Structure of the Dissertation

The dissertation consists of introduction, five chapters and conclusions. Dissertation includes 102 pages, 20 figures, 9 tables and one annex.

Chapter 1 reviews most of the work done in the field of influence of language on coding, by introducing CELP based codes, and previous researches about the topic in hand. As well, it concludes in formulating of main objective and tasks of present investigation.

Chapter 2 introduces the quality measuring techniques categorized by subjective and objective. Moreover, it contains our primary experiment on the codecs and quality assessment techniques.

Chapter 3 presents the construction of the Lithuanian speech database. After concluding the results in Chapter 2, we found ourselves in need of a well formed speech database for Lithuanian language to carry out our experiments on.

Chapter 4 presents the experimental study on the performance of CELP based codes under different languages.

Chapter 5 suggests a solution the codecs to perform better under different languages.

General conclusions as well as recommendations for further research summarize the present study. It is followed by an extensive list of references and a list of 3 publications by the author on the topic of the dissertation.

1

Literature Survey on Multi-Lingual Speech Coding

This Chapter reviews some of the research done on the influence of languages on speech and codecs, also presents investigation for this mater. This Chapter concludes in formulating of main objective and tasks of present investigation.

Part of the analysis presented in this chapter has been published in (Paulikas, *et al.* 2009).

1.1. CELP Voice Codecs

CELP voice codecs are based on LPC coding of voice. It starts with the assumption that a voice signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds (sibilants and plosive sounds) (Andersen, *et al.* 2002). Although apparently crude, this model is actually a close approximation to the reality of speech production:

- The glottis produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch);
- The vocal tract forms the tube, which is characterized by its resonances formants;

- Hisses and pops are generated by the action of the tongue, lips and throat during sibilants and plosives.

LPC analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modelled signal is called the residue. The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else.

LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter, and run the source through the filter, resulting in speech.

The more advanced CELP voice coding algorithm is based on four main ideas, (Schroeder and Atal, 1984):

- Using the source-filter model of speech production through LPC;
- Using an adaptive and a fixed codebook as the input (excitation) of the LPC model;
- Performing a search in closed-loop in a "perceptually weighted domain";
- Applying vector quantization (VQ).

The excitation is produced by summing the contributions from an adaptive (pitch) codebook and a fixed (innovation) codebook. The fixed codebook is a vector quantization dictionary that is (implicitly or explicitly) hard-coded into the codec. The entries in the adaptive codebook consist of delayed versions of the excitation. This makes it possible to efficiently code periodic signals, such as voiced sounds. The filter that shapes the excitation is called the prediction filter or source filter and is obtained using linear prediction.

In order to achieve real-time encoding using limited computing resources, the CELP search is broken down into smaller, more manageable, sequential searches using a simple perceptual weighting function.

Typically, the encoding is performed in the following order:

- LPC coefficients are computed and quantized;
- The adaptive codebook is searched and its contribution removed;
- Finally, the fixed codebook is searched.

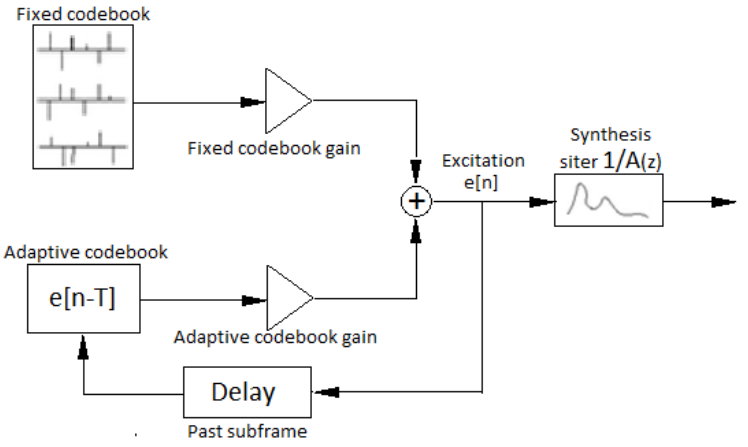

Figure 1.1. LPC model

1.1.1. AMR Codec

The AMR speech coder consists of the multi-rate speech coder, a source controlled rate scheme including a voice activity detector and a comfort noise generation system, and an error concealment mechanism to combat the effects of transmission errors and lost packets.

The multi-rate speech coder is a single integrated speech codec with eight source rates from 4.75 Kbit/s to 12.2 Kbit/s, and a low rate background noise encoding mode. The speech coder is capable of switching its bit-rate every 20 ms speech frame upon command (Uvliden, *et al.* 1998).

The audio parts including analogue to digital and digital to analogue conversion are included, to show the complete speech path between the audio input/output in the User Equipment (UE) and the digital interface of the network. The detailed specification of the audio parts is not within the scope of the present document. These aspects are only considered to the extent that the performance of the audio parts affects the performance of the speech transcoder. The speech encoder takes its input as a 13-bit uniform Pulse Code Modulated (PCM) signal either from the audio part of the UE or on the network side, from the Public Switched Telephone Network (PSTN) via an 8-bit A-law or μ-law to 13-bit uniform PCM conversion. The encoded speech at the output of the speech encoder is packetized and delivered to the network interface. In the receive direction, the inverse operations take place.

The detailed mapping between input blocks of 160 speech samples in 13-bit uniform PCM format to encoded blocks (in which the number of bits depends on the presently used codec mode) and from these to output blocks of 160

reconstructed speech samples is described in (Anskaitis, 2009). The coding scheme is Multi-Rate Algebraic Code Excited Linear Prediction.

1.1.2. Speex Codec

Unlike many other speech codecs, Speex (Valin, 2004) is not targeted at cell phones but rather at voice over IP and file-based compression. The design goals have been to make a codec that would be optimized for high quality speech and low bit rate. To achieve this, the codec uses multiple bit rates, and supports ultra wideband (32 kHz sampling rate), wideband (16 kHz sampling rate) and narrowband (telephone quality, 8 kHz sampling rate).

Speex is robust to lost packets, but not to corrupt ones since User Datagram Protocol ensures that packets either arrive unaltered or don't arrive. All this led to the choice of Code Excited Linear Prediction (CELP) as the encoding technique to use for Speex. One of the main reasons is that CELP has long proved that it could do the job and scale well to both low bit rates and high bit rates.

Decoder

Before exploring the complex encoding process of Speex we introduce the Speex decoder here. For the sake of simplicity only the narrowband decoder is presented. Figure 1.1 describes a generic CELP decoder, the excitation is produced by summing the contributions from an adaptive (*aka* pitch) codebook and a fixed (*aka* innovation) codebook:

$$e[n] = e_a[n] + e_f[n] \tag{1.1}$$

Where $e_a[n]$ is the adaptive codebook contribution, and $e_f[n]$ is the fixed codebook contribution.

The filter that shapes the excitation has an all-pole (infinite impulse response) model of the form $1/A(z)$, where $A(z)$ is called the prediction filter and is obtained using linear prediction (Levinson-Durbin algorithm). An all-pole filter is used because it is a good representation of the human vocal tract and because it is easy to compute.

In Speex, frames are 20 ms long, which is 160 samples for narrowband. Furthermore, each frame is divided into 4 sub-frames of 40 samples that are encoded sequentially.

In most modes, only the synthesis filter and the global excitation gain (see below) is encoded on a frame basis, the other parameters are encoded on a sub-frame basis.

Speex includes three main differences compared to most recent CELP codecs. First, while most CELP codecs make use of fractional pitch estimation (LeBlanc and Mahmoud, 1990) with a single gain, Speex uses an integer to encode the pitch period, but uses a 3-tap predictor (3 gains). The adaptive codebook contribution can thus be expressed as:

$$e_a[n] = g_0 e[n-T-1] + g_1 e[n-T] + g_2 e[n-T+1] \quad (1.2)$$

Where g_0, g_1, and g_2 are the pitch gains and are jointly quantized (VQ).

Many current CELP codecs use moving average (MA) prediction to encode the fixed codebook gain. This provides slightly better coding at the expense of introducing some extra dependency with past frames. Speex encodes the fixed codebook gain as a global excitation gain for the frame as well as a set of sub-frame gain corrections.

Speex uses sub-vector quantization of the innovation (fixed codebook) signal. Unfortunately, this is not a very efficient quantization method, but it was the best that could be used given patent restrictions. Each sub-frame is divided into sub-vectors of length ranging between 5 and 20 samples. Each sub-vector is chosen from a bit rate-dependent codebook and all sub-vectors are then concatenated. As an example, the 3.95 kbps mode uses a sub-vector size of 20 samples with 32 entries in the codebook (5 bits). This means that the innovation is encoded with 10 bits per sub-frame, or 2000 bps. On the other hand, the 18.2 kbps mode uses a sub-vector size of 5 samples with 256 entries in the codebook (8 bits), so the innovation uses 64 bits per sub-frame, or 12800 bps.

Encoder

The main principle behind CELP is called Analysis by-Synthesis (AbS) and means that the encoding (analysis) is performed by perceptually optimizing the decoded (synthesis) signal in a closed loop. In theory, the best CELP stream would be produced by trying all possible bit combinations and selecting the one that produces the best-sounding decoded signal. This is obviously not possible in practice for two reasons: the required complexity is beyond any currently available hardware and the "best sounding" selection criterion implies a human listener.

In order to achieve real-time encoding using limited computing resources, the CELP optimization is broken down into smaller, more manageable, sequential searches using a simple perceptual weighting function. In the case of Speex, the optimization is performed in four steps

1) Linear prediction analysis is used to determine the synthesis filter, which is converted to Line Spectral Pair (LSP) coefficients and vector-quantized.
2) The adaptive codebook entry and gain are jointly searched for the best pitch-gain combination using AbS.

3) The fixed codebook gain is determined in an "open-loop" manner only the energy of the excitation signal.
4) The fixed codebook is searched for the best entry using AbS.

Optimization for steps 2) and 4) are performed in the so called 'perceptually weighted domain', meaning that we try to minimize the perceptual difference with the original (as opposed to simply maximizing the signal-to-noise ratio). In order to do that, the following weighting filter is applied on the input signal:

$$W(z) = \frac{A\left(\frac{z}{\gamma_1}\right)}{A\left(\frac{z}{\gamma_2}\right)} \tag{1.3}$$

Where $A(z)$ is the linear prediction filter and γ_1 and $\gamma_2$control the shape of the filter by moving the poles and zeros toward the centre of the z-transform unit circle (Speex uses γ_1=0.9 and γ_2=0.6). The filter defined in 3 is in fact a very, very rough approximation for what is known as the "masking curve" in audio codecs such as Vorbis. The overall effect of the weighting filter $W(z)$ is that the encoder is allowed to introduce more noise at frequencies where the power level is high and less noise at frequencies where the power level is low.

1.2. Aspects of Multi-Lingual Speech Coding

The English language has had a dominating influence in the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust. Research carried out in paper (Parrya and Burnett, 1996) shows that coders interact with individual voices so that speech is degraded differentially for different talkers.

In the design of low-bit-rate speech coding algorithms, language variability is often considered to be of secondary importance. It is surprising to find that no comprehensive studies have ever been carried out on the role of language in speech coding. Parray, *et al.* (2000) found that quantization distortion is not uniform across languages and it influences codecs codebook performance and overall quality of decoded speech. Also a set of recommendations for codebook design for multi-lingual environments were presented there.

Other former investigations show that most low-bit-rate speech coders show bias towards non-accented English (Burnett and Parry, 1996). When the coders are used for heavily accented English or other languages, significant performance degradation is noted. An extensive study that was carried out in

(Donald, *et al.* 2002) on English, Japanese, Finish and German languages also shows domination of English language in codecs quality tests.

1.2.1. The Linguistic Sensitivities of Low Rate Prosody Modeling

All the low rate speech coders examined in this investigation use a form of pitch tracking to model the prosody of speech. Pitch tracking algorithms contribute to the minimization of low rate speech coder bandwidth use Pitch Quantization. (Parry, *et al.* 2000).

These algorithms use the correlation between pitch periods, based upon quantized thresholds, to ascertain smooth pitch contours. These thresholds, designed to detect spurious pitch periods, may not be suitable in some of the algorithm used.

Initial investigations (Itani and Paulikas, 2008) have shown some indication that these algorithms have led to an increase in pitch doubling and tripling with some forms of speech due to misjudgment of certain pitch periods and of voiced/unvoiced boundaries. Speeches displaying high pitch velocities (as is frequently the case in some accented English speech) have induced such behavior. In extreme cases some coders have even failed to regain and maintain pitch tracking.

1.2.2. The Linguistic Sensitivities of Low Rate Formant Modeling

In modeling the formant structure of speech, speech coders generally use Linear Prediction filters. The nature of these filters is described in terms of a set of coefficients known as Linear Prediction Coefficients (LPCs). Modern low rate speech coders represent the LPCs through the use of LSFs, a more robust representation of LPCs. Most other representations of LPCs are highly sensitive to small errors due to quantization or channel errors. Major bit rate reductions achieved in modern low rate speech coding are attributed in a large part to the use of quantized LSFs. Using LSFs for quantization permits a 10 coefficient LPC representation using as few as 24 bits/frame.

LSF Quantization. LSF quantization is a process whereby a finite set of LSF vectors are compared to the input LSF vector and the closest codebook vector to the input vector is used to represent that input. The codebook is generated through a process where quantized LSFs are extracted from a sequence of training speech. The nature of the training speech thus governs the nature of the LSF quantization (Burnett and Parry, 1996).

Certain speech sounds in other languages have different formant structures to that of English. As such, the process of quantization may result in the modification of their original formant structures. Thus coding may result in the modification of the true sound of the original speech.

The observations presented in this paper were only the results of an initial investigation. While in no way conclusive they do illustrate the following points;

1. The LSF clustering distributions are markedly different across the language groups represented.
2. The process of LSF quantization clearly modifies the distributions of accented speech.
3. Significant differences exist in pitch contours between accents and languages.

With a greater understanding and knowledge of the LSF behavior, across pertinent linguistic groups, new linguistically tuned codebooks could be trained and integrated to make higher performance adaptive speech coders. LSFs may prove to be a very attractive approach to accent classification. An LSF based recognition system may be an appropriate means of achieving this due to the existence of distinct linguistic LSF patterns and the inherent efficiency of employing LSFs.

1.2.3. MOS Test for Non-English Languages

Due to the "best-effort" nature of most nowadays communication (even voice communication) systems, so it is important to new how some technical solutions, e.g. voice codecs, implemented in a such system effect transferred speech quality.

In study (Tiago *et.al.* 2009) is detailed investigated voice the degradation conditions available in simulated wireless-VoIP, and conventional VoIP. The source speech signals used in that study were in English and French (four signals per language) and have been artificially corrupted to simulate distortions present in modern wireless-VoIP connections. Degradation sources that are commonly present in the wireless communications chain can include signal-based distortions such as acoustic background noise or noise suppression artifacts. These impairments are combined with distortions present in the VoIP chain, which may include codec distortions and packet loss concealment (PLC) artifacts. Speech samples were processed by three speech codecs: G.711, G.729, and Adaptive Multi rate (AMR).

In order to find out language influence on voice degradation two subjective MOS tests (one per language) were conducted. Statistics for the subjective scores collected in the listening tests are listed in Table 1.1 for the wireless-VoIP, reference, and conventional VoIP degradation conditions.

Table 1.1. Subjective score (MOS-LQS) statistics, separately for the English and French speech files, for the wireless-VoIP, reference, and conventional VoIP degradation conditions (Tiago et. al. 2009)

Statistic	Wireless-VoIP		Reference		VoIP	
	English	French	English	French	English	French
Minimum	1.05	1.02	1.30	1.11	2.52	2.09
Maximum	3.80	4.20	4.70	4.67	4.55	4.40
Average	2.43	2.42	3.77	3.63	3.67	3.36
Standard deviation	0.50	0.61	0.89	0.91	0.45	0.5

The summarized data presented in Table 1 show that there is noticeable difference in MOS quality scores between English and French languages. It can be seen that in VoIP conditions English language dominates over French language in the sense of transcoded voice quality. In wireless VoIP quality degraded for both languages is almost by same amount of MOS score but more drastically. It can be explained that in wireless conditions most influence on speech quality makes wireless channel (noise and packet loss) and language influence is masked.

In study (ISO/IEC JTC1/SC29/WG11/N2277 MPEG-4 Audio verification tests specifications - speech part, July 1998), is presented report on the MPEG-4 speech codec verification tests. The MPEG-4 Audio coding tools under test cover a bit rate range from 2 kbit/s to 64 kbit/s with a corresponding subjective audio quality. Tested codecs in narrow band mode were: CELP (Mode VIII multi rate), CELP (Mode VIII scaleable), Ref. ITU-T G723.1, Ref. ITU-T G729, Ref. GSM-EFR.

Therefore, the MPEG-4 verification tests were carried out in several parts. The tests were related first of all Internet audio applications applying codecs with bit-rates ranging from 20 to 56 kbit/s, digital audio broadcasting on AM modulated bands with bit-rates of 16 to 24 kbit/s and speech applications. This document presents the MPEG-4 audio verification test results on speech coders. The performance of speech coders is evaluated in comparison with other standard coders. In this document the results of three independent test sites are presented.

In these tests were employed speech samples of English, German, Finnish, Swedish and Japanese languages. Results obtained by NOKIA show that CELP based codecs working in narrow band mode on average got 3.02 MOS points for English, 2.87 MOS points for German and even 3.29 MOS points for Swedish

languages. NTT test results show that Japanese language often scores about 0.5 MOS point less then English.

These results show clear dependence of CELP based voice codecs performance in sense of decoded speech quality on language.

1.3. Conclusions of Chapter 1 and Formulating Tasks for the Dissertation

Based on the literature survey performed:

1. Certain speech sounds in other languages have different formant structures to that of English. As such, the process of quantization may result in the modification of their original formant structures. Thus CELP coding may result in the modification of the true sound of the original speech.
2. Speech codec verification tests shows clear dependence of CELP based voice codecs performance in sense of decoded speech quality on language.
3. Tests of the influence of language on communication systems performance were performed only using speech samples of major world languages leaving aside minor languages such us Lithuanian or Arabic languages.

In order to clarify what influences the CELP voice codecs performance in the sense of quality of decoded speech the following basic issues should be solved:

1. Investigation of speech quality measures suitable for evaluation of quality of trans-coded speech, because of there are several possibilities to evaluate speech quality. Some quality measures are based on signal properties, some of them on human perception abilities.
2. Creation of the speech records of Lithuanian language database according to ITU-T Rec. P.50 App. 1 suitable for tests of performance voice codecs, because of there is no database of Lithuanian speech records suitable for evaluation of the quality of voice communication systems.
3. Investigation of performance AMR and Speex codecs for coding non-English speech. Choice of codecs was influenced by two reasons: wide usage and approachability in order to make adjustments in codecs code.

4. Proposing and proving ways for improvement of multilingual speech coding. As so far influence of Lithuanian and Arabic languages on CELP codecs quality wasn't investigated yet, we will investigate the techniques for improving AMR and CELP codecs performance over mentioned languages.

2

Evaluation of Speech Quality

This Chapter investigates speech quality assessment techniques. These techniques revealed a defect with our primary experiment, and led us to the creation of Lithuanian speech database.

This Chapter also deals with the initial experiment with done with random speech records.

Results presented in this chapter have been published by author in (Itani *et al.* 2008, Paulikas *et al.* 2009).

2.1. Speech Quality Measures

The quality of speech signal is evaluated using objective and subjective measures. The objective measures estimates quality of speech signal by comparing its properties with non-distorted reference signal, and do not takes into account the nature of speech and perceptual abilities of a human listener.

Subjective testing is the most widely used method of assessing the quality of speech transformed by digital codecs. As a matter of fact, listening-only tests are the only feasible method of subjective testing when the transmission path is digital and/or non-linear, because of simple objective measurements are insufficient to ensure adequate transmission performance. It accounts humans

psychology and physiology (Kurasova *et. al,* 2007) aspects in speech quality evaluation. The aim of a subjective testing methodology is to measure the degradation contributed by the non-linear part of the transmission path, and hence to ensure that the performance of the complete system is satisfactory.

In the case when listening tests can't be performed or their setup is too expensive, the quality of transformed speech signals can be estimated using objective methods such as 3SQM or PESQ.

2.1.1. Objective Measures

More common objective measures are signal to noise ratio (SNR) and spectral distortions of restored signal.

A logarithmic spectral distortion for the same duration of speech signal could be calculated as follows:

$$D = \sqrt{\frac{1}{L}\sum_{0}^{L-1} 10\, log_{10}\left(\frac{S\left(e^{\frac{j2\pi k}{L}}\right)}{\hat{S}\left(e^{\frac{j2\pi k}{L}}\right)}\right)} \tag{2.1}$$

Here $S\left(e^{\frac{j2\pi k}{L}}\right)$ and $\hat{S}\left(e^{\frac{j2\pi k}{L}}\right)$ are the original and distorted speech.

A noise introduced during speech signal transformation process in voice codecs, can be estimated as ratio of signal and noise energies:

$$SNR = 10\ \log_{10}\frac{\sum_{0}^{L-1} s[n]^2}{\sum_{0}^{L-1}(s[n]-\hat{s}[n])^2} \tag{2.2}$$

Here $s[n]$ – original and $\hat{s}[n]$– distorted signals.

2.1.2. Subjective Measures

Contrarily, the subjective quality measures of speech signal encounter the perceptual abilities of a human. The subjective quality of speech usually is evaluated by group of listeners performing listening test.

The ITU-T Recommendation P.830 describes methods and procedures for conducting subjective performance evaluations of digital speech codecs. Subjective testing is also the most widely used method of assessing the quality of speech transformed by digital codecs. When the transmission path is digital and/or non-linear, simple objective measurements, such as those specified in Recommendation G.712, are insufficient to ensure adequate transmission performance. The aim of a subjective testing methodology is to measure the

degradation contributed by the non-linear part of the transmission path, and hence to ensure that the performance of the complete system is satisfactory. To be suitable for this purpose, the measurements must be:

a) reliable; and

b) carried out in a way that takes account of major interactions between the non-linear part and the other parts of the transmission system.

This implies both the ability to assign a unique numerical contribution to each digital process and the ability to use this assigned contribution in conjunction with other impairments to estimate telephone connection performance. The aim of a subjective testing methodology is to measure the degradation contributed by the non-linear part of the transmission path, and hence to ensure that the performance of the complete system is satisfactory.

A variety of methods are possible to characterize the performance of digital processes. At the present time these comprise listening-only tests involving:

a) opinion (category) ratings;

b) pair or multiple comparisons; and

c) articulation tests.

For most applications the ITU-T recommends the use of the Absolute Category Rating (ACR) method using the Listening Quality scale. However, there are times when other scales and rating methods are more suitable and appropriate (e.g. the Listening Effort scale has been found useful when interest centres on good correlation with measures of conversational performance), and these are used as well in this Recommendation. Only where there is a deviation from the use of the ACR method using the Listening Quality scale will it be stated.

The result of listening test is the average result of quality estimates of all listeners (MOS stands for Mean Opinion Score). The ITU scale is:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

When performing a subjective evaluation of a speech codec, consideration must be given to the intended application of the codec under test. The system used for recording speech materials, the selection of experimental test conditions, and the system used for delivery of processed speech materials to the listeners must be chosen so that they capture the important end-to-end characteristics of the telephone connection. In particular, the characteristics of the recording and receiving systems should reflect the characteristics of the transmitter and receiver that will be used in the intended application of the codec.

Performing subjective evaluations of digital codecs proceeds via a number of steps:

a) Preparation of source speech materials, including recording of talkers;
b) Selection of experimental parameters to exercise the features of the codec that are of interest;
c) Design of the experiment;
d) Selection of a test procedure and conduct of the experiment;
e) Analysis of results.

In the case when listening tests can't be performed or their setup is too expensive the quality of transformed speech signals can be estimated using objective methods such as 3SQM or PESQ.

The 3SQM and PESQ algorithms cannot be used to replace subjective testing but they can be applied for measurements where auditory tests would be too expensive or not applicable at all.

2.1.3. 3SQM algorithm

The 3SQM algorithm is applicable for speech quality predictions without a separate reference signal (ITU-T Recommendation P.563. 2004). This method is recommended for non-intrusive speech quality assessment, live network monitoring and assessment by using unknown speech sources at the far-end side of a connection. The 3SQM approach is the first recommended method for single-ended non-intrusive measurement applications that takes into account the full range of distortions occurring in public switched telephone networks and that is able to predict the speech quality on a perception-based scale. The calculated score is then comparable to the quality perceived by a human listener, who is listening with a conventional shaped handset at this point.

It should also be noted that the 3SQM algorithm does not provide a comprehensive evaluation of transmission quality. It only measures the effects of one-way speech distortion and noise on speech quality in the same way as it can be investigated by an auditory test assessing listening quality.

The 3SQM is a non-intrusive assessment of voice quality, as known today, can be based on two fundamentally different principles.

The first principle is looking at the signal processing to which the voice signal was exposed during the transmission, and makes assumptions on the amount of distortions introduced by the processing. The voice signal itself is not taken into account. Generally this type of algorithm can only be used with a priori knowledge of the exact transmission path and all equipment that is used in between the two endpoints of a communication path. As soon as heterogeneous networks are used, a call has to pass through foreign transit networks or the call routing is unknown, this type of assessment will fail. Frequently, also special equipment is required which traces the signal processing in routers, switches etc.

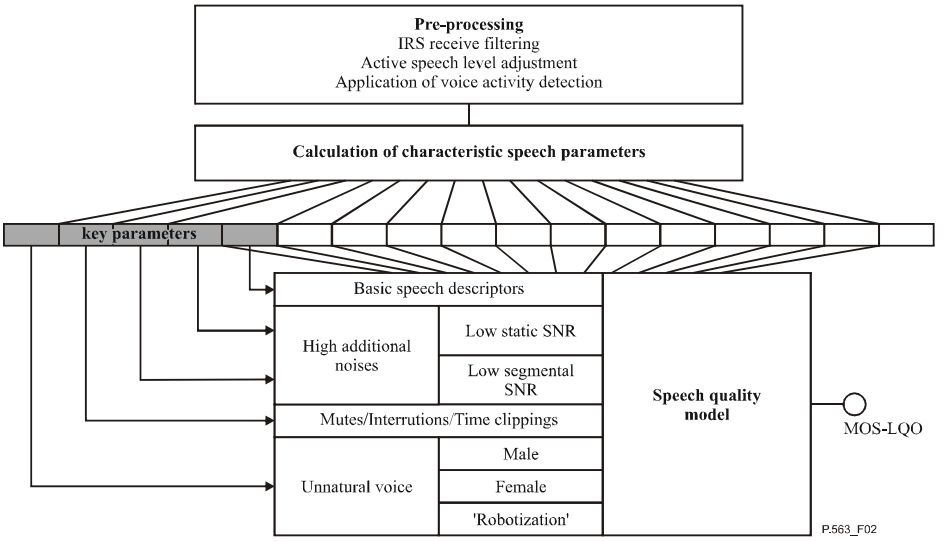

Figure 2.1. Block scheme of P.563 (ITU-T Recommendation P.563, 2004)

Such measures are currently proposed for standardization for the assessment of pure VoIP networks. However the advantage of such metrics is that they are computationally slightly less expensive than other methods. Typical examples for such algorithms are VQMon and PsyVoIP (ITU-T Recommendation P.830, 1998).

The second approach is much more universal, since in contrary to the aforementioned metrics it analyzes the voice stream and not the transmission path. Here it is possible to assess any kind of voice signal without restrictions on the network or equipment type used. Such measures are applicable in any scenario, whether the call routing is known or unknown and independent from the signal processing used. Also, no modification of existing switches etc. is required if such a metric shall be deployed, since the only required information is the speech signal itself which is available at any point in the network. Also, such metrics do not make any assumptions on the amount of distortion introduced by the network. Moreover they measure the audibility of such distortions. Measures following this approach are typically built on very general models of the human vocal tract to model the speech generation, as well as psychoacoustic models to simulate the human hearing process. These measures are – though still very efficient - slightly more complex than those relying on protocol information only, but far more flexible in their applicability. In today's

heterogeneous networks this is the only type of non-intrusive measurement that can be used with hardly any restrictions.

The P.563 approach could be visualized as an expert who is listening to a real call with a test device like a conventional handset into the line in parallel. This visualization explains also the main application and allows the user to rate the scores gained by P.563. The quality score predicted by P.563 is related to the perceived quality by linking a conventional handset at the measuring point.

Consequently, the listening device has to be part of the P.563 approach. Therefore, each signal will first be pre-processed. This pre-processing begins with the model of the receiving handset. Following this, a voice activity detector (VAD) is used to identify portions of the signal that contain speech and the speech level is calculated. Finally, a speech level adjustment to –26 dBov is applied. The pre-processed speech signal to be assessed will be investigated by several separate analyses, which detect, like a sensor layer, a set of characterizing signal parameters. This analysis will be applied at first to all signals. Based on a restricted set of key parameters, an assignment to a main distortion class will be made.

The key parameters and the assigned distortion class are used for the adjustment of the speech quality model. This provides a perceptual based weighting where several distortions are occurring

ITU-T Rec. P.563 (05/2004) in one signal but one distortion class is more prominent than the others. The basic block-scheme of P.563 is shown in Figure 2.1.

Before the voice signal can be assessed properly it needs to be preprocessed in a first step. The important steps of preprocessing are:

- IRS receive filtering: The employed filter simulates a standard handset used in the laboratories for the subjective listening tests.
- Speech level adjustment.
- Separation in voice and non-voice parts via Voice Activity Detection (VAD).

In a second stage the distortion and speech parameters are extracted for the speech signal. They are divided up into three main functional blocks which also correspond to the in recommendation P.563 considered main distortion classes. The main distortion classes are defined as:

1. Vocal tract analysis and unnaturalness of speech
   - Basic speech quality depending on whether the talker is male or female.
   - Robotic voice, e.g. caused by band limitation in GSM networks and unnatural voice like beeps.
2. Analysis of strong additional noise
   - Low static SNR (Background noise floor).
   - Low segmental SNR (Noise that is related to the signal's envelope).

3. Interruptions, mutes and time clipping
    - Impairments as a result of lost packets in packet based transmission systems.

All of these classes are based on very general principles which make no assumptions on the underlying network or distortion types occurring under certain conditions. The only prerequisite is the scientific knowledge on how human speech is generated and how it is perceived by human beings. This knowledge is built into the distortion model and does not vary with the application.

During the workings for the standardization of P.563 the developers found, that several output parameters can be clustered to define single isolated distortion classes (see previous subsection). This models the phenomenon that any human listener focuses on the foreground of the signal stream. That is the listener would not judge the quality of the transmitted voice by a simple sum of all occurred distortions but because of a single dominant noise artifact in the signal.

Those distortion classes can be identified from a subset of the extracted parameters and are then prioritized according to the distortion's relevance with respect to the average listeners' opinions.

The dominant distortion classes used with 3SQM™ are:

- Low static SNR: Occurs with a high background noise level.
- Mutes: Loss of packets in packet based transmission systems.
- Low segmental SNR.
- Unnatural voice – Robotization: Highly periodic signal due to band limitation e.g. in GSM networks.
- Basic speech quality: In case, if the other models do not apply. Here two different models are used depending on whether the talker is male or female.

This part of the algorithm models the cognitive feature of human perception.

Furthermore, simple methods for delay identification do not perform satisfactorily when the system under test includes coding distortions and filtering. Matters are further complicated by the fact that a growing number of communications systems may introduce variable delay, in particular due to packet-based transmission.

The ITU-T has now standardized a new model that was required to show good performance for a very wide range of applications, both in codec assessment and also end-to-end testing of networks of all types.

With low bit-rate coding, errors, and the potential for delay to vary during a test, time alignment is a difficult problem. Common 'classical' techniques, such as transfer function estimation or cross-correlation, are not robust and can give biased, or highly inaccurate, estimates in many cases. The new techniques for

delay identification that have been introduced in this paper work well across a large database of simulated and measured network conditions including many different systems and types of delay change.

The piecewise constant delay assumption appears to be valid for many applications, including common variable delay communications systems such as VoIP.

The final speech quality is calculated by combining the intermediate quality result with some additional signal features, as shown in Figure 2.2.

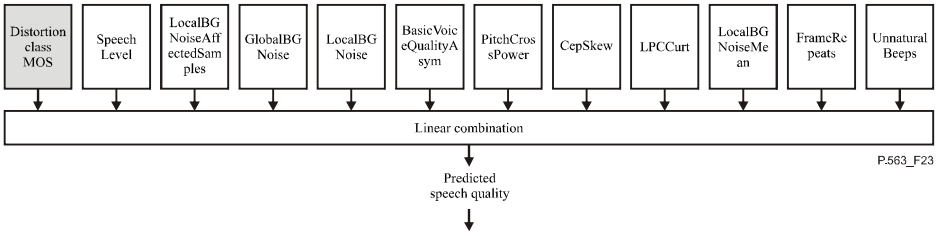

Figure 2.2. Overall speech quality prediction (ITU-T Recommendation P.563, 2004)

2.1.4. PESQ algorithm

PESQ (Perceptual Evaluation of Speech Quality) compares an original signal with a degraded signal that is the result of passing original signal through a communications system (Joskowicz, *et al.* 2009). The output of PESQ is a prediction of the perceived quality that would be given to degraded signal by subjects in a subjective listening test. PESQ compares the original input signal with the aligned degraded output of the device under test using a perceptual model (ITU-T Recommendation P.862, 2001) (**Figure 2.3**).

In the first step of PESQ a series of delays between original input and degraded output are computed. For each of these intervals a corresponding starts and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. Based on the set of delays that are found, PESQ compares the original input signal with the aligned degraded output of the device under test using a perceptual model.

The key to this process is transformation of both the original and degraded signals to an internal representation that is analogous to the psychophysical

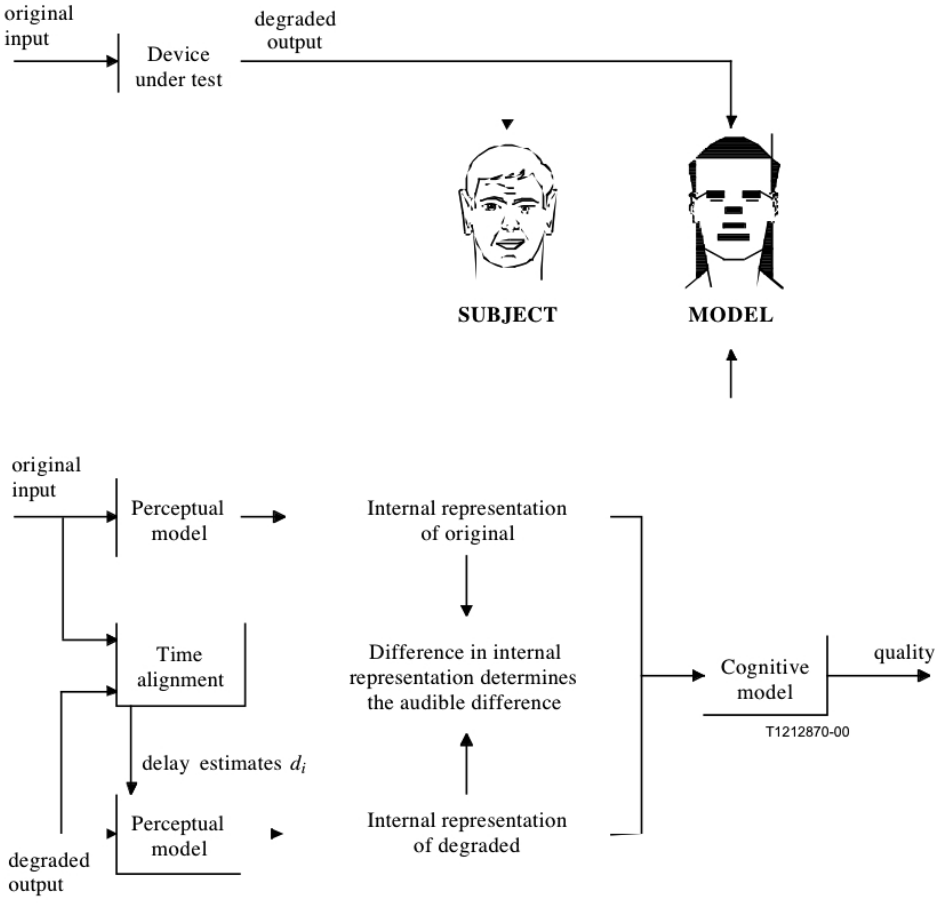

Figure 2.3. Overview of the basic philosophy used in PESQ (ITU-T Recommendation P.862, 2001)

representation of audio signals in the human auditory system, taking account of perceptual frequency and loudness.

This is achieved in several stages: time alignment, level alignment to a calibrated listening level, time frequency mapping, frequency warping, and compressive loudness scaling. The internal representation is processed to take

account of effects such as local gain variations and linear filtering that may have little perceptual significance.

More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. This allows a small number of quality indicators to be used to model all subjective effects. In PESQ, two error parameters are computed in the cognitive model; these are combined to give an objective listening quality.

During the development of PESQ it was found that introducing time alignment techniques made a significant improvement to the model's accuracy, especially with variable delay conditions. This sets PESQ apart from earlier models, such as PSQM and MNB, that did not include time alignment processes and were therefore unsuitable for use in end-to-end measurement applications.

A key limitation of early perceptual models for assessing the subjective quality of codecs (Joskowicz, J. *et al.* 2009) is that they do not include a method of identifying delay. This makes them unsuitable for end-to-end measurement applications, in which the delay is unknown.

Furthermore, simple methods for delay identification do not perform satisfactorily when the system under test includes coding distortions and filtering. Matters are further complicated by the fact that a growing number of communications systems may introduce variable delay, in particular due to packet-based transmission.

In contrast, PESQ was evaluated against a demanding set of conditions including variable delay, filtering, coding and errors – sometimes all at the same time. The results presented in the accompanying paper (ITU-T Recommendation P.862) illustrate that PESQ gives accurate quality scores in many different applications.

The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value. The range of the PESQ score is –0.5 to 4.5, although for most cases the output range will be a listening quality MOS-like score between 1.0 and 4.5, the normal range of MOS values found in an ACR experiment.

2.2. Pilot Analysis of Language Influence on Codecs Performance

This experiment was done on a poorly constructed speech database that does not follow the structure described in the ITU-T recommendation P.50.

2.2.1. Experiment Setup

For the experiment we took 60 speech samples (24 English, 20 Arabic, and 16 Lithuanian). All the samples were 16 bit PCM formats sampled at 8 kHz wave files. Then we encoded and decoded these samples with two codecs AMR and Speex. To evaluate the coding quality we used PESQ and 3SQM algorithms.

SQM (we used ITU-T P.563 2004 and ITU-T P.862 2001) and PESQ packages contains the C code files and it alouds to compile it on various platforms.

All the speakers were native, but they were all males. Moreover, the words and the phrases used did not include all the language alphabet and sounds.

2.2.2. Experiment Results

Experiment results are depicted in Figures 2.1 and 2.2. Figure 2.1 shows 3SQM quality scores of all 60 samples, while Figure 2.2 presents the PESQ scores.

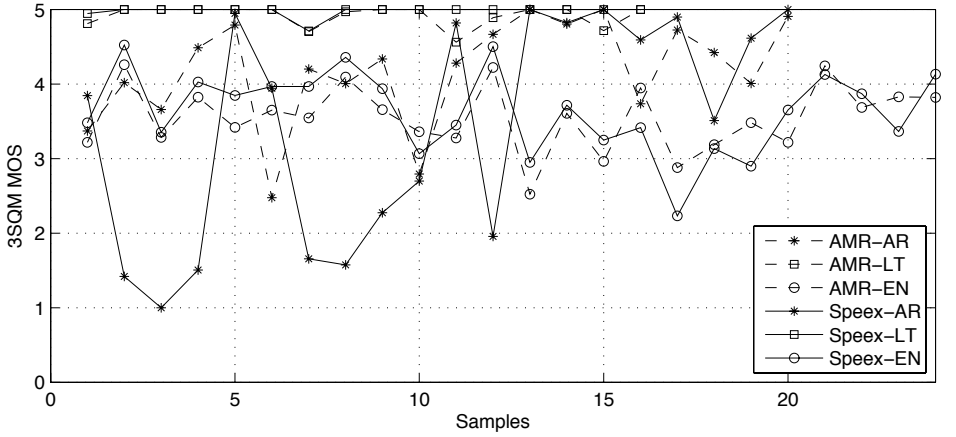

Figure 2.4. 3SQM results for Arabic (AR), Lithuanian (LT) and English (EN) languages

Table 2.1. Mean values and standard deviations of 3SQM results from **Figure 2.4**

MOS	AMR			Speex		
	AR	LT	EN	AR	LT	EN
Mean	4.18	4.91	3.54	3.45	4.97	3.63
Std.	0.70	0.14	0.44	1.50	0.07	0.55

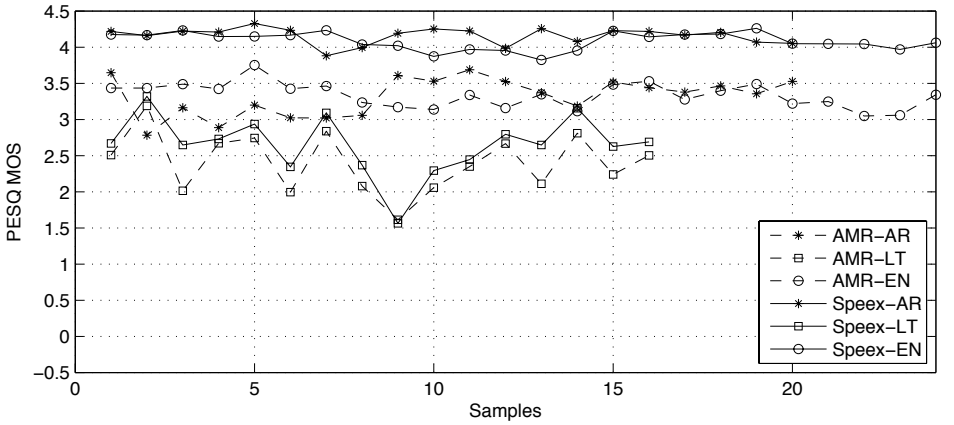

Figure 2.5. PESQ results for Arabic (AR), Lithuanian (LT) and English (EN) languages

Table 2.2. Mean values and standard deviations of PESQ results from **Figure 2.5**

MOS	AMR			Speex		
	AR	**LT**	**EN**	**AR**	**LT**	**EN**
Mean	3.31	2.39	3.33	4.15	2.64	4.08
Std.	0.26	0.41	0.17	0.11	0.41	0.12

The mean values and the standard deviations of experimental data for 3SQM are summarized in Table 2.1. Where, Table2.2 shows the mean values and the standard deviations for PESQ.

From the presented 3SQM results (Fig. 2.1 and Table 2.1), it can be seen that 3SQM algorithm erroneously performs on samples of Lithuanian language. It can be explained that duration utterances of Lithuanian language are too short for 3SQM thus these results are inadequate. In case of samples of Arabic and English languages the quality scores are spread over wide interval but results produced by AMR and Speex codecs are related. Both codecs perform more stable on English language.

Analyzing PESQ results, the correlation between performances of both codecs can be observed. Results are more stable, because of PESQ method for quality estimation uses undistorted reference signal thus these results are more reliable than 3SQM results. In summary, both investigated codecs show the quality bias towards non-accented English.

2.3. Conclusions of Chapter 2

The results presented showed that objective quality measures are not enough. The human listening always plays a big role, when it comes to quality, after all these codecs are made to be used by people. Thus, subjective quality measures should be carried out.

The results of the pilot investigation of the voice codecs performance did not show a clear weakness in the performance of CELP based codecs toward a certain language. The mean values were close. But the standard deviation shows that the codecs were performing in a more stable way with English rather than Arabic and Lithuanian.

The observations presented here are only the results of an initial investigation. While in no way conclusive, they illustrate some tendencies of language influence to quality variation of CELP codecs.

However, in order to clearly show what influences different quality results for different languages – codecs or maybe quality estimation algorithms—an additional more extensive analysis should be performed.

3

Creation of Speech Records Database of Lithuanian Language

The experiment results of the pilot analysis of language influence on AMR and Speex codecs performance show some weakness in the performance of CELP based codecs toward non-English language. The mean values were close. But the standard deviation shows that the codecs were performing in a more stable way with English rather than Arabic and Lithuanian.

The observations presented in this paper are only the results of an initial investigation. While in no way conclusive, they illustrate some tendencies of language influence to quality variation of CELP codecs. However, in order to clearly show influence on coded speech quality results for different languages an additional more extensive analysis should be performed and more reliable speech material must be used.

As speech material from Appendix I of ITU-T Rec. P.50 that is dedicated for assessment of voice codecs quality does not include speech records of Lithuanian language, the appropriate speech records database should be recorded and tested to conform requirements of ITU-T Rec. P.50.

Results presented in this chapter have been published by author in (Paulikas *et al.* 2010) and in report of scientific project "*Investigation of Lithuanian Language Influence on Coding of Speech signals*".

3.1. Existing Databases of Speech Records of Lithuanian Language

Investigations related with processing of speech of Lithuanian language have been carried out long time ago. Considerable work put into this area is by scientists from the Institute of Mathematics and Informatics, Vytautas Magnus University and the Kaunas Technological University (Simutis *et. al.*, 2009, Maskeliunas *et. al.*, 2009). However, works of scientist from these institutions are mostly related with speech recognition. Specific features of speech signal for recognition purpose are extracted using traditional speech processing methods that were designed and tested for English language.

In telecommunications while designing voice codecs it is very important to encounter aspects of different languages. For tests of speech communication systems it is recommended to use speech records from database of Appendix I of ITU-T Rec. P.50. This database contains speech records from 20 different languages, however, records of Lithuanian language are absent.

It is need to mention, that in Lithuania have been created several speech records databases of Lithuanian language by: Institute of Mathematics and Informatics, Lithuanian Radio News Agency (LRN0 and LRN0.1), Vytautas Magnus University (VDU-RTG) and Kaunas Technological University (LTDIGITS).

3.1.1. LRN Database

There is two versions of LRN database: LRN0 and its extension LRN0.1 (Laurinčiukaitė *et al.*, 2006). Database LRN0 contains ~10 hours duration records of continuous Lithuanian speech. These records are taken directly from Lithuanian Radio 1 broadcasted news records. Each record is around 12 minutes duration, 11 kHz discretization and 16 bits quantization. Records are of high quality with clear and correct pronunciations. LRN0 contains 141 records (~18000 words) of 23 announcers (males and females) individually segmented in sciences together with written text.

LRN0.1 database is of 17 hours (also 18000 words) of 31 announcers (males and females) duration with word level annotations. However, around 80 % of all records belong to ten announcers (6 males and 4 females).

Both database are of the form of file records that are separated by extension and name. Extension *.wav denotes speech record, extension *.txt – transcription of record. File name is of the form: [science_id].[extention]. Here [science_id] is of the form: [RZ][S][YY][MM][DD]_- [H]_[OR], here RZ – constant, S – announcer ID, YY – year, MM – month, DD – day, H – hour and OR – science number in original record.

LRN0.1 database is constantly expanded by adding records of 3–4 hours duration each year. The purpose of database is speech recognition applications.

3.1.2. VDU-RTG Database

This database contains records of continuous speech by professional speakers (male and female) (Raškinis *et al.,* 2008). Total duration of records is around 17.5 hours of 114130 (among them 33645 different) words. Texts are chosen from Lithuanian newspapers (~3 hours), popular periodic literature (~3 hours) and works of fiction (~11.5 hours). Each speech record is around 13 minutes of duration. All records are made in silent environment and stored in 41 kHz and 16 bits PCM format *.wav files. In total, database consists of 258 files.

VDU-RTG database is denoted for speech recognition experiments.

3.1.3. LTDIGITS Database

This database is created in Kaunas Technical University under supervision of A. Rudžionis (Rudžionis, 2001). The the database is targeted for applications such as recognition of Lithuanian digits, control command recognition, phoneme recognition and discrimination. Database consists of 9600 words records pronounced by 200 speakers (100 male and 100 female).

3.2. Databases of Speech Records of Lithuanian Language Compatible with ITU-T Rec. P.50 Appendix I

3.2.1. Characteristics of ITU-T Rec. P.50 Appendix I

The speech signals used in the pilot experiment were poor in quality, they were recorded using laptop computer with built in microphone. Moreover, they do not represent a statistically good sample of the whole language. This made us looks for batter samples for our future experiments, which made us use the samples from ITU-T recommendation P.50 Appendix I.

This Appendix to Recommendation P.50 is a CD-ROM containing useful test signals for telephone transmission quality, telephone installations, local line networks. The signals on this CD-ROM include the signal described in Recommendation P.50 as well as other signals that have been found useful by some Administrations. Additionally, the full speech database that was used to develop Recommendation P.50 is also on this CD-ROM.

Speech database consists of records of 20 languages and accents: English (American) and (British), Arabic, Chinese (Mandarin), Danish, Dutch, Finnish, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Norwegian, Polish, Portuguese (Brazilian), Russian, Spanish (Castilian), Swedish. Each language (accent) is represented by 16 records (8 – male and 8 – female). Each record, of duration about 6–8 s, consists of 2–3 phrases. Records are written as 16 bits mono PCM files with 16 kHz sample rate. File names are in the form: Lt_f1.wav, where Lt – represents language (Lt – Lithuanian), f or m – female or male, correspondingly and 1 – record number.

3.2.2. The ITU-T Recommendation P.50 Samples

In the following experiments we will use the speech records samples of American and British English, Arabic and Lithuanian Languages. For the American English language samples in the ITU-T Recommendation P.50 are organized as depicted in Table 3.1.

Table 3.1. American English language samples in the ITU-T Recommendation P.50

Language: American English (AT&T)	
File Name	Sentence
A_eng_m1	The ship was torn apart on the sharp reef. Sickness kept him home the third week. The box will hold seven gifts at once. Jazz and swing fans like fast music.
A_eng_m2	
A_eng_m3	
A_eng_m4	
A_eng_f1	
A_eng_f2	
A_eng_f3	
A_eng_f4	
A_eng_f5	A rod is used to catch pink salmon. The source of the huge river is the clear spring. Kick the ball straight and follow through. Help the woman get back to her feet.
A_eng_f6	A siege will crack the strong defense. Grape juice and water mix well. Roads are paved with sticky tar. Fake stones shine but cost little.

A_eng_m5	The birch canoe slid on the smooth planks. Glue the sheet to the dark blue background. It's easy to tell the depth of a well. Four hours of steady work faced us.
A_eng_m6	Rice is often served in round bowls. The young kid jumped the rusty gate. Guess the results from the first scores. A salt pickle tastes fine with ham.
A_eng_f7	High seats are best for football fans. Tea served from the brown jug is tasty. A dash of pepper spoils beef stew. A zestful food is the hot-cross bun.
A_eng_f8	Heave the line over the port aide. A lathe cuts and trims any wood. It's a dense crowd in two distinct ways. His hip struck the knee of the next player.
A_eng_m7	The swan dive was far short of perfect. The beauty of the view stunned the young boy. Two blue fish swam in the tank. Her purse was full of useless trash.
A_eng_m8	A speedy man can beat this track mark. He broke a new shoelace that day. The coffee stand is too high for the couch. The urge to write short stories is rare.

For the British English language samples in the ITU-T Recommendation P.50 are organized as depicted in Table 3.2.

Table 3.2. British English language samples in the ITU-T Recommendation P.50

Language: British English	
File Name	Sentence
B_eng_f1	I was away for nine weeks. The dining-room was lit by gas. There were no vegetables left.
B_eng_f2	It's human nature to blame another. He had completely forgotten his hat. I did not wish him to know.
B_eng_f3	The act was a deliberate murder. He was attracted by her face. The timber fell across the road.
B_eng_f4	We were compelled to stay there. He had a curious sense of power. This car appears to have six wheels.
B_eng_f5	This is no business of yours. What did he say he went up for?

	She was in bed when she got home.
B_eng_f6	The king had a navy at sea. I have told you everything. Have you the letter with you?
B_eng_f7	He came running up to us. You haven't got the words right. My cup of trouble is now full.
B_eng_f8	The store was open on Saturday. She always asked him what to do. Now they had some packing to do.
B_eng_m1	I think it would be wonderful. There will be guests coming. She seldom listened to anybody.
B_eng_m2	He could not remember his name. I never can leave you two alone. I shall entreat his pardon.
B_eng_m3	They have a good idea of fairness. You must go and do it at once. Have you said all you have to say?
B_eng_m4	He was not in the mood for music. I can well understand your feelings. I have rented a small house.
B_eng_m5	Not a flower bloomed in the garden. I have never had time to learn. She was not impressed by his talk.
B_eng_m6	We poured cold water on his head. There was not much time left. It is useless waiting for a taxi.
B_eng_m7	He looked about him again. They met again in the station. Why must there be all this secrecy?
B_eng_m8	I saw it with my own eyes. The note was immediately dispatched. He wanted to leave college.

Arabic language samples in the ITU-T Recommendation P.50 are organized as depicted in Table 3.3.

Table 3.3. Arabic language samples in the ITU-T Recommendation P.50

Language: Arabic	
File Name	Sentence
Ar_f1	**تعتبر الجمعية العامة جهاز الأمم المتحدة الرئيسي لأجراء المداولات.**
Ar_f2	**يقع مقر منظمة الأمم المتحدة في نيويورك.**

Ar_f3	رب انك وهبتني نجاحا فلا تنزع تواضعي.
Ar_f4	رأيت النفس تكره ما لديها و تطلب كل ممتنع عليها.
Ar_f5	و أخيرا اتفق الأعضاء على ارجاء الاجتماع الى أجل غير مسمى.
Ar_f6	ليس البر أن تولوا وجوهكم قبل المشرق و المغرب.
Ar_f7	و ليل كموج البحر أرخى سدوله علي.
Ar_f8	دخل الرجل البيت ولم يخرج منه.
Ar_m1	و قل رب ارحمهما كما ربياني صغيرا.
Ar_m2	على قدر أهل العزم تأتي العزائم.
Ar_m3	كل من تلقاه يشكو دهره ليت شعري هذه الدنيا لمن؟
Ar_m4	الأمم المتحدة هي اتحاد منظم بين دول مستقلة ذات سيادة.
Ar_m5	أود أولا و قبل كل شيئ أن أرحب بكم جميعا.
Ar_m6	لا يخفى عليكم أن الهدف من اجتماعنا هو تبادل الآراء بشأن السلم العالمي.
Ar_m7	يؤدي الترجمون الفوريون دورا هاما في الاضطلاع بمهمة الاتصالات الشاقة.
Ar_m8	جنيف مدينة رائعة, لكن تنقصها الرمال الدافئة.

As speech samples of Lithuanian language are not included ITU-T Rec. P.50 Appendix I database, we will record them to be similar to given examples of other languages.

3.2.3. Lithuanian Speech Records Database

The Lithuanian speech records database should not only satisfy requirements of ITU-T Rec. P.50 Appendix I, but also present the specifics of common Lithuanian language. At first, sentences should rep- resent most frequently used language parts. Second, sentences must consist of most frequently used words.

According to research done in (Žilinskienė, 2002) and (Žilinskienė, 2003), the most frequent language parts in Lithuanian texts are nouns (~45 %), verbs (~20 %) and adjectives or pronouns (~8 %). The most frequently used nouns are the following: *galva, laikas, darbas, metas, vanduo, gyvenimas, motina, miestas, vakaras, širdis, pasaulis, saulė, dangus, vėjas;* verbs: *būi, galėti, nebūti, turėti, žinoti, eiti, sakyti, norėti, reikėti, matyti, žiūrėti, pasakyti, negalėti, imti, gyventi, atrodyti, ateiti, suprasti, stovėti, sėdėti, kalbėti;* and adjectives: *aukštas, valstybinis, administracinis, pagrindinis, politinis, bendras, naujas, didelis, tarptautinis, konstitucinis, ekonominis, atskiras, socialinis, visuomeninis.*

Basing on this research, four sentences consisting of 2–3 phrases were composed and recorded:

1. *Jonas buvo jaunas vyras, kuris moterims atrodė gražus, bet jo keistos kalbos jos negalėjo suprasti.*
2. *Taip galėjo sakyti tik žmogus, kurio širdis žinojo, kad šiame gyvenime galima pasitikėti tik savo tėvu.*
3. *Vakare aplink jį susirinkė stovėjo vyrai, tarytum atėję iš tuščio pasaulio, kuriame nebuvo saulės ir dangaus.*
4. *Po sunkaus darbo mieste motinai skaudėjo galvą, tačiau jos veidas visada buvo gražus, o balsas gyvas*.

Lithuanian language samples according to the ITU-T Recommendation P.50 were organized as depicted in Table 3.4.

Table 3.4. Lithuanian language samples in the ITU-T Recommendation P.50

Language: Lithuanian	
File Name	Sentence
Lt_f1	*Jonas buvo jaunas vyras, kuris moterims atrodė gražus, bet jo keistos kalbos jos negalėjo suprasti.*
Lt_f2	
Lt_m1	
Lt_m2	
Lt_f3	*Taip galėjo sakyti tik žmogus, kurio širdis žinojo, kad šiame gyvenime galima pasitikėti tik savo tėvu.*
Lt_f4	
Lt_m3	
Lt_m4	
Lt_f5	*Vakare aplink jį susirinkė stovėjo vyrai, tarytum atėję iš tuščio pasaulio, kuriame nebuvo saulės ir dangaus.*
Lt_f6	
Lt_m5	
Lt_m6	
Lt_f7	*Po sunkaus darbo mieste motinai skaudėjo galvą, tačiau jos veidas visada buvo gražus, o balsas gyvas.*
Lt_f8	
Lt_m7	
Lt_m8	

3.3. Verification of Database of Speech Records of Lithuanian Language

In order to verify quality of recorded samples, we employed 3SQM quality estimation algorithm that is applicable for speech quality predictions without a separate reference signal (ITU-T Rec. P.563). We tested two Lithuanian database versions: the candidate (in this article denoted as LT1) recorded using high quality recording equipment and microphone, and the reference database (here denoted as LT2) recorded with built in laptop sound input device. LT2 has significantly lower signal to noise ratio then LT1. For comparison purposes, the same test was carried out on American English, British English and Arabic languages records taken from ITU-T Rec. P.50 Appendix I database. Obtained results (Fig. 3.1) show that individual and average quality estimates of records of all tested languages vary in a wide range. So 3SQM is not appropriate for absolute quality rating, however in the case of Lithuanian language we can state that LT1 version is a better candidate than LT2.

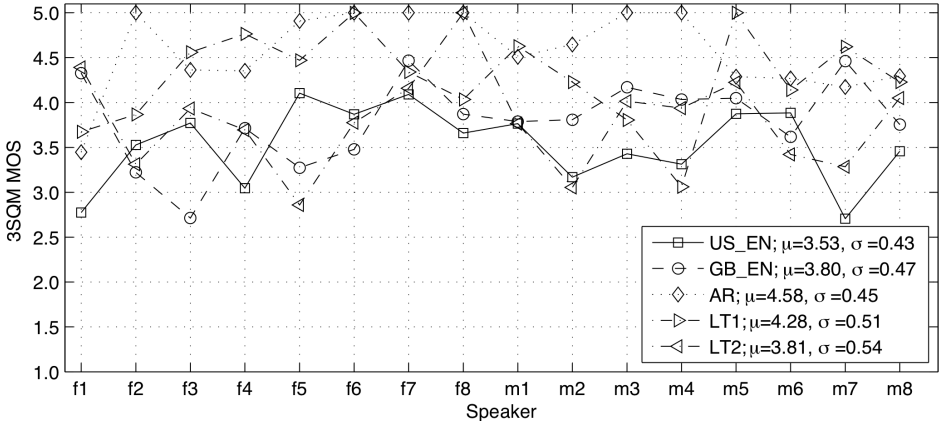

Figure 3.1. 3SQM scores for proposed Lithuanian (LT1 and LT2) records database as well as American (US), British (GB) English, Arabic (AR) speech records from ITU-T Rec. P.50 Appendix I database

Further, to prove the applicability of our proposed Lithuanian speech records database for evaluating quality of voice communication systems we experimentally tested the quality of two most popular voice codecs, AMR and Speex, which are widely used in telecommunications.

3.4. Conclusions of Chapter 3

Basing on the review of existing Lithuanian speech records databases, Lithuanian language usage statistics analysis and according to requirements of ITU-T Rec. P.50 Appendix I the database of Lithuanian speech records was created and tested.

Quality test results show that In the case of Lithuanian language we can state that LT1 version is a better candidate than LT2, because of higher quality scores.

Obtained quality test results show that individual and average quality estimates of records of all tested languages vary in a wide range. So 3SQM is not appropriate for absolute quality rating.

Further, to prove the applicability of our proposed Lithuanian speech records database for evaluating quality of voice communication systems we experimentally tested the quality of two most popular voice codecs, AMR and Speex, which are widely used in telecommunications.

4

Investigation of Codecs Performance in Multi-Lingual Speech Coding

This Chapter investigates the performance of speech codecs in a multi-lingual environment. It shows the setup and the results of an experiment, conducted with two codes in three different languages. This Chapter also deals with quality measurements, subjective and objective.

Results presented in this chapter have been published by author in (Paulikas *et al.* 2010).

4.1. Experiment Setup

In order to judge the quality of AMR and Speex codecs in sense of transformed voice signal (coded and decoded) firstly we employed two objective quality measures–SNR and spectral distortions (SD) calculated according to Eq. (2.1) and Eq. (2.2). Then we obtain the PESQ mean opinion score as discussed in Chapter 2. We are using three quality measurement techniques, to ensure conclusive results.

The speech signals used in this experiment were taken from the ITU–T Recommendation P.50 Appendix I. We used from this database the speech signals for United States English, Great Britain English, and Arabic. Whereas,

the speech signals for the Lithuanian language were taken from the database that we constructed and described in Chapter 3.

The speech signals were coded and decoded using two codecs, AMR and Speex. The AMR codec is distributed as C language pre-compiled binary library that lets to write and build a simple program for coding and decoding voice records. However modification of operation algorithms is not possible.

On the other hand, the Speex codec (we used Version 1.2 Beta 3) comes as open source project with freely distributed C language code. That allows modifying the employed operation algorithms.

In both codecs cases as input signal we used speech records stored in *.wav (PCM) file format. The output of both codecs is also the *.wav file.

The coding and decoding procedure is done for Speex codec as:

```
hostname$speexenc -dtx MODE wavFileName tmp.amr
hostname$speexdec tmp.amr tmp.wav
```

and for AMR codec as:

```
hostname$sox wavFileName -r 8000 tmp.sw
hostname$amrnb-encoder test.sw tmp.amr
hostname$amrenc -dtx MODE tmp.sw tmp.amr
hostname$amrnb-decoder tmp.amr tmp.sw
hostname$sox -r 8000 tmp.sw tmp.wav
```

here `sox` command is used to strip wav file header as AMR codec works only with plain sound samples and MODE determines average codec's bitrate: for AMR codec it is set to one of:

```
MODE = {'MR475'; 'MR515'; ' MR59'; ' MR67'; ' MR74';
'MR795'; 'MR102'; 'MR122'}
```

and for Speex:

```
MODE = {4750; 5150; 5900; 6700; 7400; 7950; 10200; 12200}.
```

Then the quality of the new signals again was measured by calculating SNR and spectral distortion (Eqs. 2.1 and 2.2), and using PESQ algorithm as shown below:

```
hostname$pesq +8000 wavFileName tmp.wav
```

All code used in these experiments is presented Annex.

4.2. Results of Tests of Objective Quality of Coded Multi-Lingual Speech Signals

Performed experiment results are shows that in Figures 4.1 and 4.2.

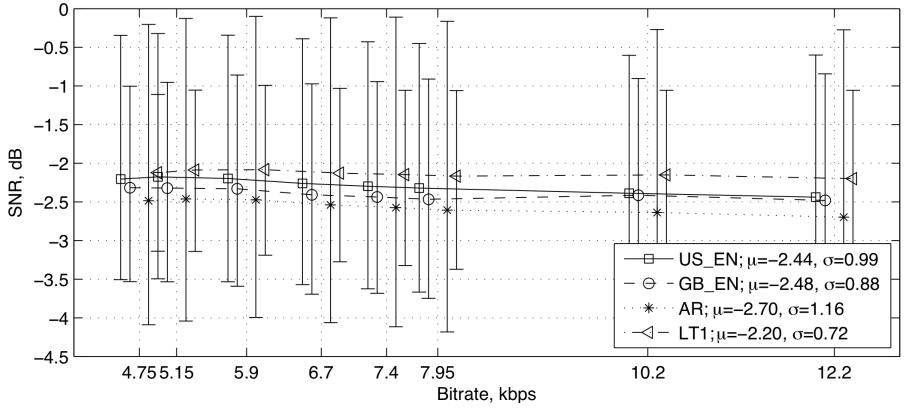

Figure 4.1. Mean and min and max values of SNR for AMR codec

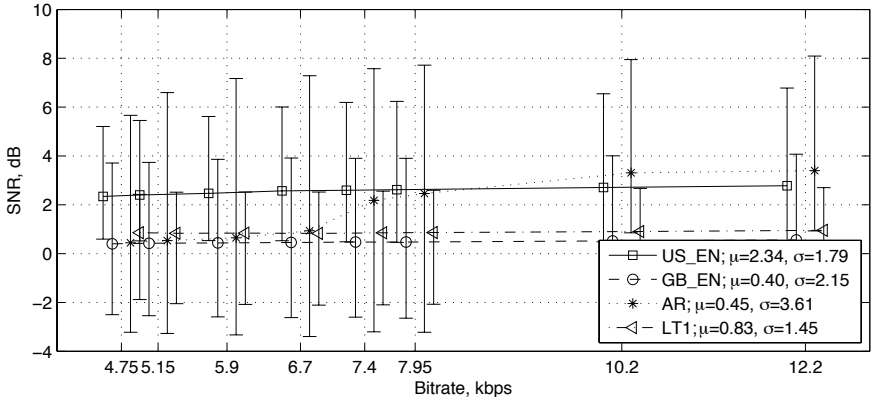

Figure 4.2. Mean and min and max values of SNR for Speex codec

Depicted SNR measure is not appropriate for evaluation of quality of transformed speech signal. Mean values remains almost constant while dispersion is very big. It can be explained as just a small shift in time between original and transformed speech signals introduces big difference between samples, however according to human perception a small delay (shift in time) does not introduces any degradation of speech.

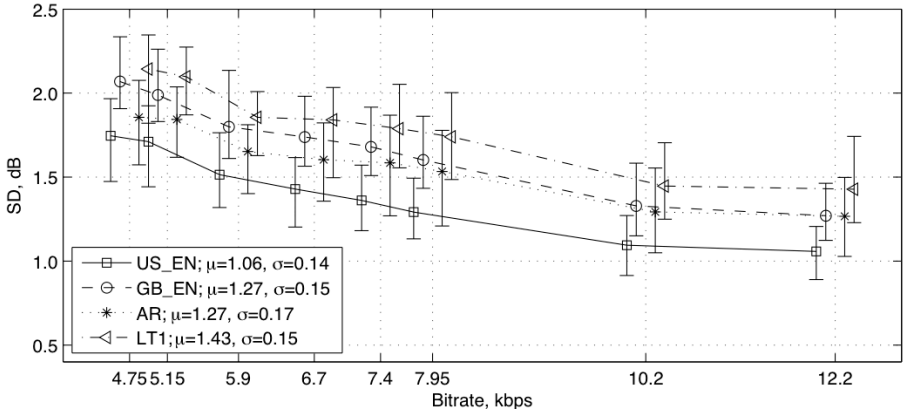

Figure 4.3. Mean and min and max values of spectral distortions for AMR codec

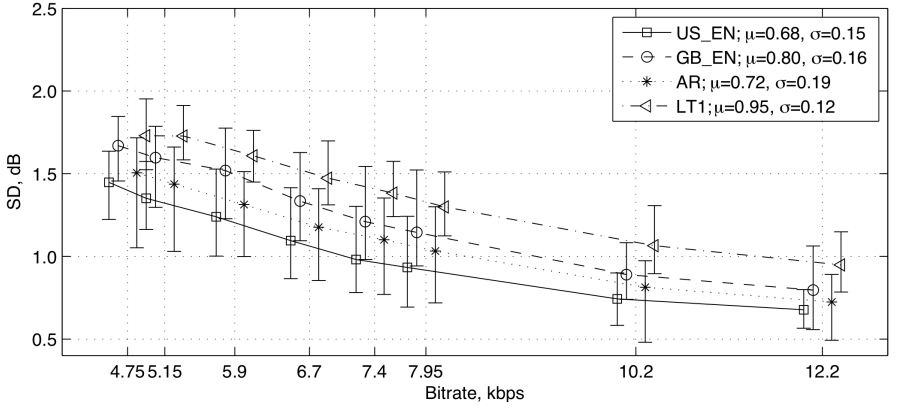

Figure 4.4. Mean and min and max values of spectral distortions for Speex codec

The spectral distortions (see Figs. 4.3 and 4.4) eliminates effect of delay, and are more suitable for judging the quality of transformed speech signal and related with it performance in sense of reproduced speech signal quality of voice codecs. As it can be seen, by increasing the bitrate spectral, distortions decrease for both codecs. Also, distortions for speech samples of non-English languages in general are bigger and/or spread in wide interval.

To assess codecs quality in the sense of human perception we used the PESQ algorithm. It compares the original input signal with the aligned degraded output of the device under test using a perceptual model. A PESQ result is an objective MOS, which is mapped to the MOS scale that directly expresses the voice quality. The PESQ MOS as defined by the ITU recommendation P.862

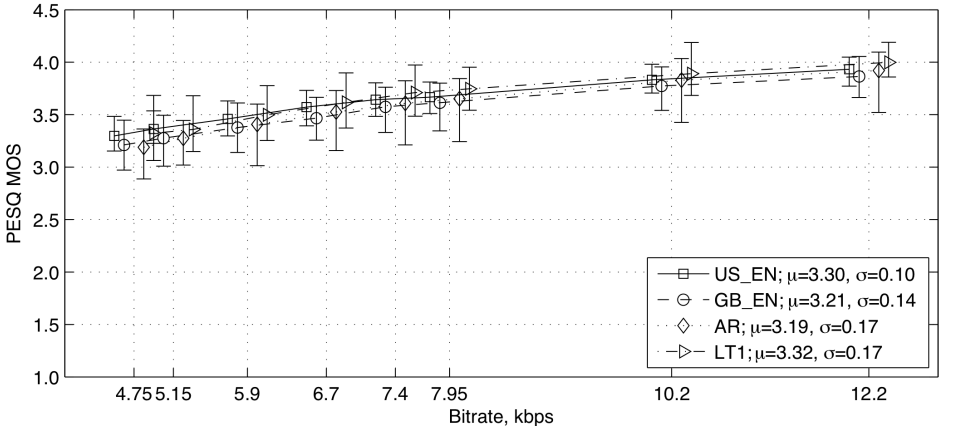

Figure 4.5. Mean and min and max values of PESQ MOS scores for AMR codec

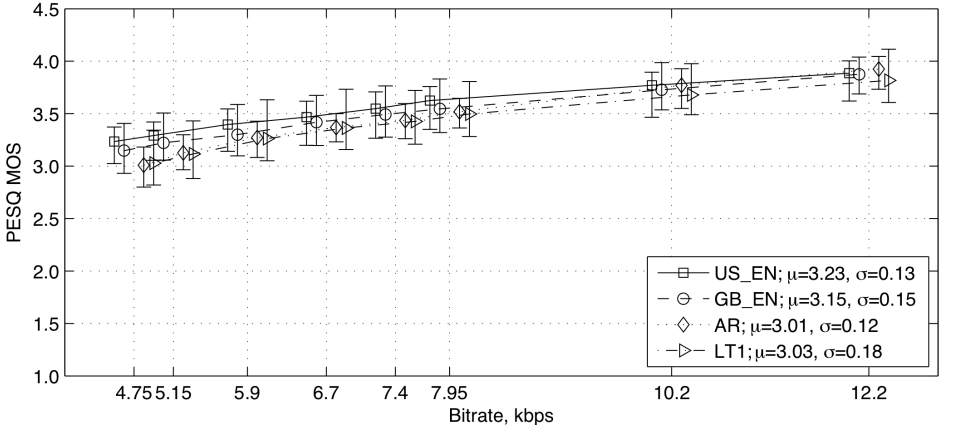

Figure 4.6. Mean and min and max values of PESQ MOS scores for Speex codec

ranges from 1.0 (worst) up to 4.5 (best). The ITU scale ranges up to 5.0, but PESQ simulates a listening test and is optimized to reproduce the average result of all listeners. Statistics however prove that the best average result one can generally expect from a listening test is not 5.0 (”excellent”) but 4.5.

For the following experiment we employed both composed Lithuanian speech records database and English (American and British) and Arabic speech records database from ITU-T Rec. P.50 Appendix 1. During tests codecs were used in narrow band mode and set to operate at various bitrates. As bitrates of AMR codec are fixed, we also setup Speex codec to work at same bitrates as AMR codec. The quality of encoded and decoded speech was estimated using

PESQ algorithm. MOS scores for AMR and Speex codecs are depicted in Figs. 4.5 and 4.6, respectively.

From Fig. 4.5 can be seen that average quality estimates for speech records of all tested languages are close. Difference in the range of 0.1 MOS points is statistically insignificant. From Fig. 4.6 it is noticeable that Speex codec performed at lower quality than AMR on all tested languages. Also Speex codec shows less stability at lower bitrates i.e. difference in MOS points greater than 0.5.

Furthermore, as was already noted in [6] both codecs per- forms more stable on English language and in Speex codec's case it exhibits about 0.2 greater MOS values.

4.3. Subjective MOS Tests of Quality of Coded Multi-Lingual Speech Signals

To make sure that obtained codecs' performance results for speech sample of Lithuanian and Arabic languages are correct, we performed subjective listening test by auditory from six persons whose native spoken language is Lithuanian.

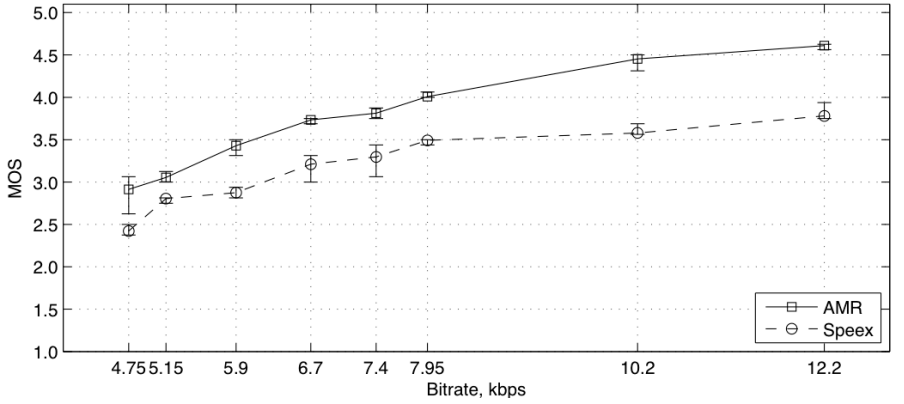

Figure 4.7. MOS scores for AMR and Speex codecs obtained during listening test

During this test, all listeners individually estimated quality (in MOS scale) of degraded speech record by comparing it with the original one. Quality estimates from all six listeners were averaged to produce MOS. Obtained results are shown in Fig. 4.7. This test also confirms that AMR codec performs better than Speex and quality estimates have very close tendency to the one obtained by using PESQ algorithm.

4.4. Conclusions of Chapter 4

Language has influence on performance of tested voice codecs. Their stability of quality of reproduced speech significantly decreases at lower bitrates when coding speech signals of non-English languages. Furthermore, there also exists a noticeable bias to quality of reproduced speech of English language.

Experiments show that both codecs performs more stable on English language and in Speex codec's case it exhibits about 0.2 greater MOS values.
In narrow band mode, AMR codec exhibits better and more stable performance (in the sense of quality of reproduced speech) than open source Speex codec.

The listening tests results also confirms that AMR codec performs better than Speex and quality estimates have very close tendency to the one obtained by using PESQ algorithm.

5

Improving of Speex Codecs Performance

This Chapter considers CELP codebook design in order to adapt it for samples of particular language. Further in this Chapter the results of Speex codec codebook adaptation for Lithuanian and Arabic languages and obtained increase performance are presented and discussed.

Results presented in this chapter have been published by author in the report of scientific project "*Investigation of Lithuanian Language Influence on Coding of Speech signals*" and presented in international conference "The 9th IEEE International Symposium on Signal Processing and Information Technology. December 14–17, 2009 – Ajman, UAE".

5.1. Code Book Adaptation for Speech Samples of Specific Language

5.1.1. CELP Fundamentals

Consider the CELP analysis structure shown in Figure 5.l. The long term (quantized) inverse filter (with $2q + 1$ non-zero taps), $B(z)$, for subframe n is given by:

$$B(z) = 1 - \sum_{k=-q}^{q} b_k z^{-(M+k)} \tag{5.1}$$

and the short term (quantized) inverse filter (order p), $A(z)$, for subframe n is given by:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{5.2}$$

The perceptual weighting filter, which attempts to obtain a larger signal to noise ratio in inter-formant regions is given by:

$$W(z) = \frac{A_p(z/\beta)}{A_p(z/\gamma)} \tag{5.3}$$

where γ and β are optimized based on subjective measures, and $A_p(z)$ is the optimum unquantized inverse filter (for subframe n).

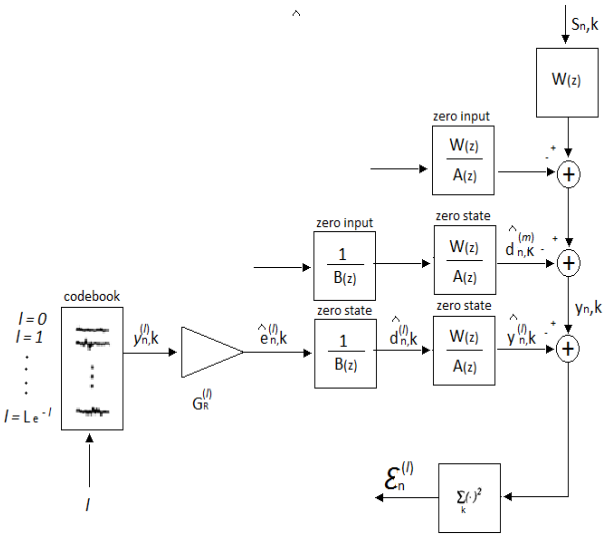

Figure 5.1. CELP search procedure. The codebook dimension, or subframe size is $\boldsymbol{K}_C$. The index n is over all subframes, and the index k is over all elements of a particular subframe. Thus, $s_{n,k}$ is the k-th element of the n-th subframe (LeBlanc *et. al.* 1990).

Typically, $A(z)$ is determined to minimize the open loop residual energy, and $B(z)$ is determined (closed loop) to minimize the noise weighted error before determination of the fixed codebook (further in this chapter it will be referred as codebook) excitation (the energy in $y_{n,k}$). The determination of these parameters and complexity reduction techniques based on the structured codebooks is referred in (Davidson *et. al.* 1989). Overlap is often used to reduce block coding edge effects. That is, components of the excitation vector near the end of a subframe have little effect on the current subframe, but may adversely affect future subframes. Overlap considers the influence of these elements by letting the filters ring for K_O samples after the last sample in the excitation vector.

The weighted mean squared error for a particular codebook index l over a subframe (at subframe index n) of dimension K_C with overlap K_O is given by (LeBlanc *et. al.* 1990):

$$\epsilon_n^{(l)} = \left|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n^{(l)}\right|^2 = \boldsymbol{y}_n^T\boldsymbol{y}_n - 2G_n^{(l)}\boldsymbol{y}_n^T\boldsymbol{H}_n\boldsymbol{v}^{(l)} + G_n^{(l)2}\boldsymbol{v}^{(l)T}\boldsymbol{H}_n^T\boldsymbol{H}_n\boldsymbol{v}^{(l)} \tag{5.4}$$

where the $K_C + K_O$ by K_C dimensional lower triangular Toeplitz matrix $\boldsymbol{H}_n$ represents the zero state filtering operation (of $W(z)/A(z)$). The l-th excitation (column) vector $\boldsymbol{v}^{(l)}$ is of dimension K_C, and the (column) vectors $\boldsymbol{y}_n$ and $\hat{\boldsymbol{y}}_n^{(l)}$ are of dimension $K = K_C + K_O$.

The codebook design algorithms are all based on the Generalized Lloyd Algorithm (GLA) (Machoul *et. al.* 1985) and require a sufficiently rich training sequence to design the codebook. Due to the long memory in $1/B(z)$, the algorithm is not guaranteed to converge to a local minimum. That is, the set of training vectors $\mathcal{T} = \{\boldsymbol{y}_n\}$ changes from one iteration to the next. The problem arises because (for simplicity) we assume the training vectors do not depend on the codevectors. Due to the long and short term predictor memory this is not the case. In practice, convergence is similar to the GLA, although the average weighted mean squared error has been observed to increase (slightly) after some iterations.

The optimum codebook is defined as that which minimizes Eq. (5.4) over the whole training sequence. We minimize:

$$\bar{\epsilon}_n^{(l)} = \sum_{n=0}^{N-1} \epsilon_n^{(l)} \tag{5.5}$$

The index n is over all training vectors ($\boldsymbol{y}_n$), l_n is the optimum codebook index for training vector (or subframe) n, $\boldsymbol{v}^{(ln)}$ is the optimum codevector (for subframe n) and G_n is the optimum gain for codevector $\boldsymbol{v}^{(l_n)}$. The codebook design techniques are all based on minimization of Eq. (5.5). All design techniques assume training vector $\boldsymbol{y}_n$ is not a function of the current, or past codevectors.

In next Subsection we consider general codebook design. The codebook is given by L_C distinct K_C dimensional codevectors.

5.1.2. Codebook Design

We now discuss techniques whereby near optimal codebooks may be designed. The technique is based on a vector quantizer design algorithm using the noise weighted mean squared error distortion measure. Due to the influence of previous codevectors on future codevectors (via the long term predictor memory), only suboptimal codebooks may be designed, (the error is not guaranteed to decrease continually to a local optimum). In practice, the average distortion usually decreases until a local optimum is found, then oscillates slowly in the vicinity of that local optimum.

The design goal is to minimize Eq. (5.5) over all possible codebooks of size L_C and dimension K_C. Given a training sequence of N speech vectors $\boldsymbol{S} = \{s_n\}$, and an initial codebook $C^{(0)} = \{\boldsymbol{v}^{(l)}\}$, we analyze the vectors using the CELP structure to obtain the training set $\mathcal{T}^{(0)} = \{\boldsymbol{y}_n\}$. Essentially, we use the initial codebook to partition the training sequence $\left(\mathcal{T}^{(0)}\right)$ into L_C cells, or regions $R^{(j)}$ according to the nearest neighbor search, and compute new centroids (or codevectors) for the regions. Cell j is comprised of those subframes which have $\boldsymbol{v}^{(l_n)} = \boldsymbol{v}^{(j)}$ (the optimum codebook index at time n is j). Eq. (5.5) can then be split up into L_C terms, one term for each particular cell:

$$\bar{\epsilon}_n^{(l)} = \sum_{n \in R^{(0)}} \left|\boldsymbol{y}_n - G_n \boldsymbol{H}_n \boldsymbol{v}^{(0)}\right|^2 + \sum_{n \in R^{(1)}} \left|\boldsymbol{y}_n - G_n \boldsymbol{H}_n \boldsymbol{v}^{(1)}\right|^2 + \quad \ldots \\ + \sum_{n \in R^{(L_C-1)}} \left|\boldsymbol{y}_n - G_n \boldsymbol{H}_n \boldsymbol{v}^{(L-1)}\right|^2 \tag{5.6}$$

where the summation indicates summation over the region in which all codevectors are identical. Minimization of Eq. (5.6) is equivalent to minimizing each term, since a particular codevector only influences the summation in its region. Furthermore, in each region (j), $\boldsymbol{v}^{(l_n)}$ does not depend on n (since $j = l_n$. Thus we minimize (with respect to $\boldsymbol{v}^{(j)}$):

$$\epsilon_n^{(l)} = \sum_{n \in R^{(j)}} \left(\boldsymbol{y}_n^T \boldsymbol{y}_n - 2G_n^{(l)} \boldsymbol{y}_n^T \boldsymbol{H}_n \boldsymbol{v}^{(l)} + G_n^{(l)2} \boldsymbol{v}^{(l)T} \boldsymbol{H}_n^T \boldsymbol{H}_n \boldsymbol{v}^{(l)}\right) \tag{5.7}$$

for each region j, $0 \le j < L_C$. Since $\boldsymbol{v}^{(l_n)} = \boldsymbol{v}^{(j)}$ is a constant for each region and does not depend on the index n, we may write:

$$\epsilon_n^{(l)} = \bar{\sigma}_y^{(j)2} - 2\bar{\boldsymbol{c}}^{(j)} \boldsymbol{v}^{(j)} + \boldsymbol{v}^{(j)T} \bar{\boldsymbol{R}}^{(j)} \boldsymbol{v}^{(j)} \tag{5.8}$$

where

$$\bar{\sigma}_y^{(j)2} = \sum_{n \in R^{(j)}} \boldsymbol{y}_n^T \boldsymbol{y}_n \tag{5.9a}$$

$$\bar{\boldsymbol{c}}^{(j)} = \sum_{n \in R^{(j)}} G_n \boldsymbol{H}_n^T \boldsymbol{y}_n \tag{5.9b}$$

and

$$\bar{\boldsymbol{R}}^{(j)} = \sum_{n \in R^{(j)}} G_n^2 \boldsymbol{H}_n^T \boldsymbol{H}_n \tag{5.9c}$$

It can be shown that to minimize Eq. (5.8)

$$\boldsymbol{v}^{(j)} = \left(\bar{\boldsymbol{R}}^{(j)}\right)^{-1} \bar{\boldsymbol{c}}^{(j)} \tag{5.10}$$

Now we have a new codebook ($C^{(1)}$), which can be used in the CELP analysis structure to obtain the training set $\mathcal{T}^{(1)}$. Unlike typical VQ design techniques, the training set $\mathcal{T}^{(1)}$ will not be the same as $\mathcal{T}^{(0)}$. The above design algorithm is just a simple extension of the GLA for a CELP type distortion based on the above assumptions.

5.1.3. Speex Codec Codebook

Speex codec codebook is implemented as K_C by M dimensional matrix, where $L_C = 2^M$, M – number of bits in codebook index. Then, excitation can be given by $\boldsymbol{v}^{(j)} = \bar{C}\boldsymbol{b}^{(l)}$, where $\bar{C}$ codebook (K_C by M dimensional matrix) and $\boldsymbol{b}^{(l)}$ – l-th codeword (M dimensional column vector). Alliteratively, yet equivalently, the excitation can be written as $\boldsymbol{v}^{(l)} = \boldsymbol{B}^{(l)}C$ where C is a $K_C M$ dimensional column vector (containing the stacked columns of $\bar{C}$) and $\boldsymbol{B}^{(l)}$ is a K_C by $K_C M$ dimensional Toeplitz matrix, with the first row having elements $b_k^{(l)}$ in positions B_{0,kK_c}.

Over the training sequence, we may write:

$$\bar{\epsilon} = \sum_{n=0}^{N-1} \left(\boldsymbol{y}_n - G_n \boldsymbol{H}_n \boldsymbol{v}^{(l_n)}\right)^2 \tag{5.11}$$

Substituting $\boldsymbol{v}^{(l)} = \boldsymbol{B}^{(l)}C$ into Eq. (5.11) leads to:

$$\bar{\epsilon} = \sum_{n=0}^{N-1} \left(\boldsymbol{y}_n - G_n \boldsymbol{H}_n \boldsymbol{B}^{(l_n)} C \right)^2 = \bar{\sigma}_y^2 - 2\bar{\boldsymbol{c}}^T C + C^T \bar{\boldsymbol{R}} C \tag{5.12}$$

where

$$\bar{\sigma}_y^{(j)2} = \sum_{n=0}^{N-1} \boldsymbol{y}_n^T \boldsymbol{y}_n \tag{5.13a}$$

$$\bar{\boldsymbol{c}} = \sum_{n=0}^{N-1} G_n \boldsymbol{H}_n \boldsymbol{B}^{(l_n)T} \boldsymbol{H}_n^T \boldsymbol{y}_n \tag{5.13b}$$

and

$$\bar{R} = \sum_{n=0}^{N-1} G_n^2 \boldsymbol{B}^{(l_n)T} \boldsymbol{H}_n^T \boldsymbol{H}_n \boldsymbol{B}^{(l_n)} \tag{5.13c}$$

The Speex codebook C is computed by solving system of linear equations:

$$C = \bar{R}^{-1} \boldsymbol{c} \tag{5.14}$$

As design of totally new fixed codebook is very difficult and time consuming task, we choose to adapt existing Speex codec fixed codebook for training samples of Arabic and Lithuanian speech. We expect that trained excitation codebook can reproduce more accurate and complex excitation signals.

The modified codebook is obtained by adapting the excitation vectors to a training speech sequence by applying iterative optimization scheme based on classic LBG algorithm (Linde *et. al.* 1980). According to LBG algorithm, two steps are performed successively. I the first step, the data training set is vector coded using the given codebook (original Speex codec's fixed codebook). In the second step, the centroid of all data vector assigned to same codeword is calculated.

In order to assure a minimum coding error in CELP coding process, in the training process is used an error measure similar to the error measure used in excitation vector search process of encoder. For a given set of N-dimensional training data $\{y_i(n)\}$ the cumulated error for all input vector belonging to region P_j of a codebook vector c_j is given by:

$$\epsilon_j = \sum_i \frac{\sum_{n=0}^{N-1} \left(x_i(n) - G_i c_j(n) * h_i(n) \right)^2}{\sum_{n=0}^{N-1} x_i^2(n)} \tag{5.15}$$

where $x_i(n)$ is the i-th data vector, G_i is the optimum gain value and $h_i(n)$ denotes the impulse response of the cascaded synthesis filter (A(z), $B(z)$) and weighting filter ($W(z)$). Minimizing ϵ_j leads to a set of linear equations for a given value G_i and each iteration to find the optimum codebook is followed by an iteration in order to optimize gain value.

5.2. Modified Codebook of Speex Codec

Speex codec has six codebooks for excitation in narrowband CELP mode for 2000, 3200, 4000, 7000, 9600 and 12800 bps. For example, in codebook for 9600 bps has 64 excitation vectors with 5 samples each (Table 5.1).

Table 5.1. Codebook for excitation in narrowband CELP mode (9600 bps)

Excitation vectors
{1,5,-15,49,-66}, {-48,-4,50,-44,7}, {37,16,-18,25,-26},
{-26,-15,19,19,-27}, {-47,28,57,5,-17}, {-32,-41,68,21,-2},
{64,56,8,-16,-13}, {-26,-9,-16,11,6}, {-39,25,-19,22,-31},
{20,-45,55,-43,10}, {-16,47,-40,40,-20}, {-51,3,-17,-14,-15}
{-24,53,-20,-46,46}, {27,-68,32,3,-18}, {-5,9,-31,16,-9},
{-10,-1,-23,48,95}, {47,25,-41,-32,-3}, {15,-25,-55,36,41},
{-27,20,5,13,14}, {-22,5,2,-23,18}, {46,-15,17,-18,-34},
{-5,-8,27,-55,73}, {16,2,-1,-17,40}, {-78,33,0,2,19},
{4,53,-16,-15,-16}, {-28,-3,-13,49,8}, {-7,-29,27,-13,32},
{20,32,-61,16,14}, {41,44,40,24,20}, {7,4,48,-60,-77},
{17,-6,-48,65,-15}, 32,-30,-71,-10,-3}, {-6,10,-2,-7,-29},
{-56,67,-30,7,-5}, {86,-6,-10,0,5}, {-31,60,34,-38,-3},
{24,10,-2,30,23}, {24,-41,12,70,-43}, {15,-17,6,13,16},
{-13,8,30,-15,-8}, {5,23,-34,-98,-4}, {-13,13,-48,-31,70},
{12,31,25,24,-24}, {26,-7,33,-16,8}, {5,-11,-14,-8,-65},
{13,10,-2,-9,0}, -3,-68,5,35,7}, {0,-31,-1,-17,-9},
{-9,16,-37,-18,-1}, {69,-48,-28,22,-21}, {-11,5,49,55,23},
{-86,-36,16,2,13}, {63,-51,30,-11,13}, {24,-18,-6,14,-19},
{1,41,9,-5,27}, {-36,-44,-34,-37,-21}, {-26,31,-39,15,43},
{5,-8,29,20,-8}, {-20,-52,-28,-1,13}, {26,-34,-10,-9,27},
{-8,8,27,-66,4}, {12,-22,49,10,-77}, {32,-18,3,-38,12},
-3,-1,2,2,0}

Now, our goal is to adapt existing Speex codec fixed codebooks for training samples of Arabic and Lithuanian speech. The adaptation procedure was performed iteratively using half of speech samples from Arabic and Lithuanian speech records databases. The adaptation criteria were mean square error

betweent original and and decoded speech signals. The example of modified for Lithuanian language codebook for 9600 bps has is shown in Table 5.2.

Table 5.2. Modified codebook for Lithuanian language for excitation in narrowband CELP mode (9600 bps)

Excitation vectors
{1,5,-14,46,-61},{-45,-4,47,-41,7},{34,15,-17,23,-24},
{-24,-14,18,18,-25,},{-44,26,53,5,-16},{-30,-38,63,20,-2},
{60,52,7,-15,-12},{-24,-8,-15,10,6},{-36,23,-18,20,-29},
{19,-42,51,-40,9},{-15,44,-37,37,-19},{-47,3,-16,-13,-14},
{-22,49,-19,-43,43},{25,-63,30,3,-17},{-5,8,-29,15,-8},
{-9,-1,-21,45,88},{44,23,-38,-30,-3},{14,-23,-51,33,38},
{-25,19,5,12,13},{-20,5,2,-21,17},{43,-14,16,-17,-32},
{-5,-7,25,-51,68},{15,2,-1,-16,37},{-73,31,0,2,18},
{4,49,-15,-14,-15},{-26,-3,-12,46,7},{-7,-27,25,-12,30},
{19,30,-57,15,13},{38,41,37,22,19},{7,4,45,-56,-72},
{16,-6,-45,60,-14},{30,-28,-66,-9,-3},{-6,9,-2,-7,-27},
{-52,62,-28,7,-5},{80,-6,-9,0,5},{-29,56,32,-35,-3},
{22,9,-2,28,21},{22,-38,11,65,-40},{14,-16,6,12,15},
{-12,7,28,-14,-7},{5,21,-32,-91,-4},{-12,12,-45,-29,65},
{11,29,23,22,-22},{24,-7,31,-15,7},{5,-10,-13,-7,-60},
{12,9,-2,-8,0},{-3,-63,5,33,7},{0,-29,-1,-16,-8},
{-8,15,-34,-17,-1},{64,-45,-26,20,-20},{-10,5,46,51,21},
{-80,-33,15,2,12},{59,-47,28,-10,12},{22,-17,-6,13,-18},
{1,38,8,-5,25},{-33,-41,-32,-34,-20},{-24,29,-36,14,40},
{5,-7,27,19,-7},{-19,-48,-26,-1,12},{24,-32,-9,-8,25},
{-7,7,25,-61,4},{11,-20,46,9,-72},{30,-17,3,-35,11},
{-3,-1,2,2,0}

In order to compare original and modified codebooks we calculated the normalized Euclidean distance between vectors of original ($\boldsymbol{v}^{o}$) and modified ($\boldsymbol{v}^{m}$) codebooks:

$$d_{\boldsymbol{c}} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\sum_{i=1}^{N}\left(v_i^{o} - v_i^{m}\right)^2} \tag{5.16}$$

where M– number of vectors in codebook, N– excitation vector length.

As it can be seen from Figure 5.2, large differences (at average around 6–8 %) between codebooks excitation vectors are observed at lower bitrates. At large bitrates these differences are a lot smaller (at average around 1–3 %). It can explain large diference between Speex codec quality for coding English and Lithuanian and Arabic speech at lower bitrates.

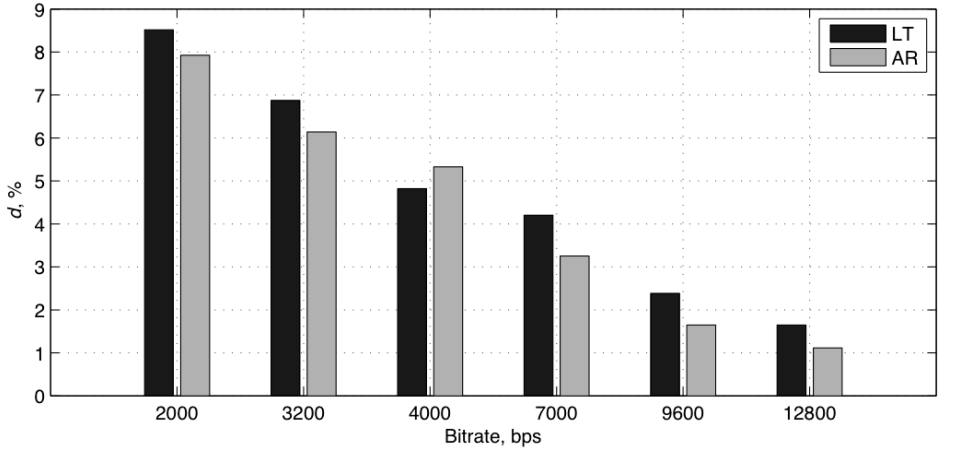

Figure 5.2. Normalized Euclidean distance between original and modified Speex codec codebooks

5.3. Evaluation of Modified Speex Codec Performance

Speex codec codebook was adapted using speech records from created Lithuanian and Arabic (taken from ITU-T Rec. P50) databases. For training of new fixed codebook we employed only half of speech records samples from each language. We took four records of male and four records of female speakers. The Speex codecs fixed codebooks were adapted for all in tests employed voice signal coding bitrates. Afterwards, performance of modified Speex codec was evaluated using all speech records samples for each language (Arabic and Lithuanian). The performance, again, was evaluated by computing spectral distortions introduced by modified Speex codec and PESQ MOS quality of encoded speech. In these experiments Speex codec version 1.2rc1 was used.

In Figures 5.2 and 5.3 presented comparison of spectral distortions and PESQ MOS, respectively, for Lithuanian speech samples between original and modified Speex codecs. In Figures 5.4 and 5.5 presented comparison of spectral distortions and PESQ MOS, respectively, for Arabic speech samples between original and modified Speex codecs.

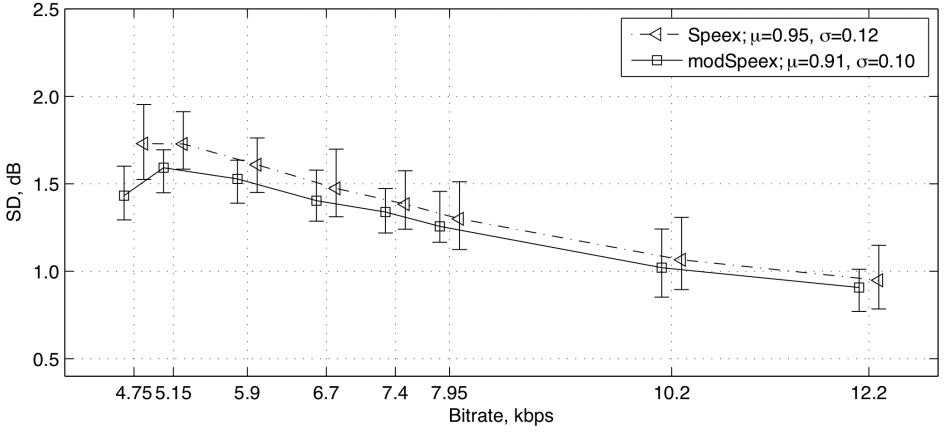

Figure 5.3. Comparison of spectral distortions introduced by original and modified Speex codecs for speech records of Lithuanian language

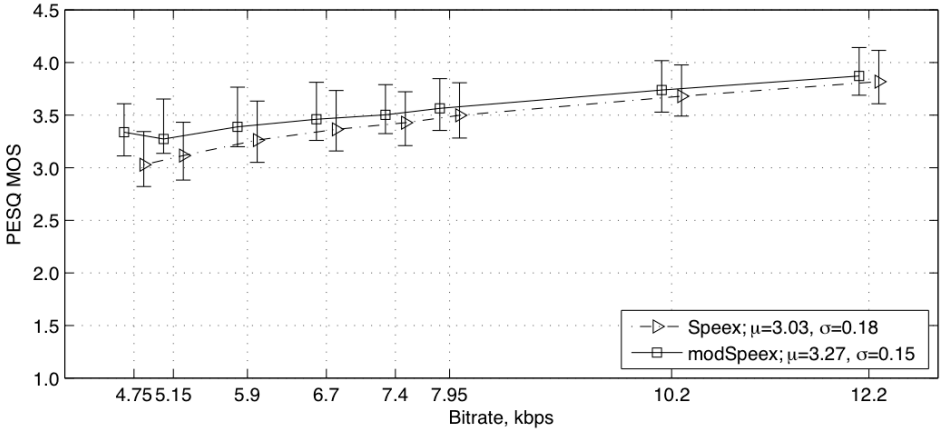

Figure 5.4. Comparison of PESQ MOS values between original and modified Speex codecs for speech records of Lithuanian language

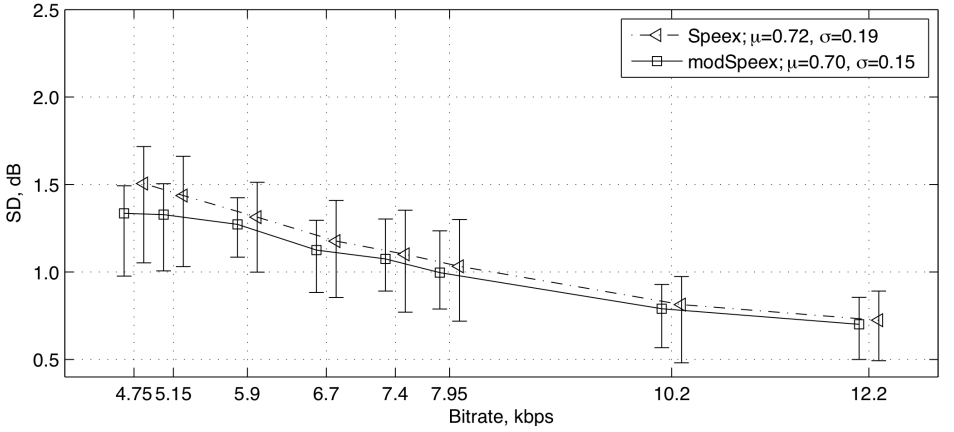

Figure 5.5. Comparison of spectral distortions introduced by original and modified Speex codecs for speech records of Arabic language

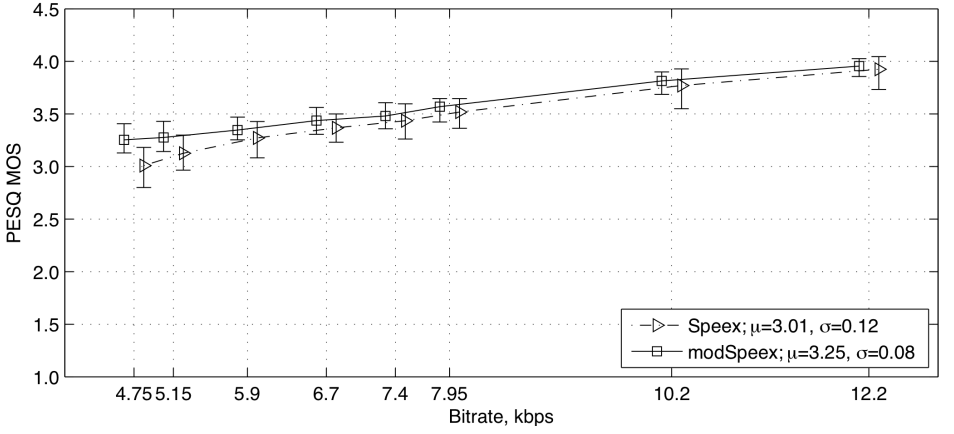

Figure 5.6. Comparison of PESQ MOS values between original and modified Speex codecs for speech records of Arabic language

Presented results show that Speex codec's codebook adaptation in general improves quality of coded speech by lowering mean values of spectral distortions and its diference between maximum and minimum values and by increasing mean PESQ MOS values and decreasing its diference between maximum and minimum values. In average MOS values increase about 0.2 points along with improving stability, narrower spread of MOS values for individual speech samples (in Figures 5.2–5.5 indicated by error bars).

The increase in Speex codec's performance is more noticeable at lower bitrates then at higher.

5.4. Conclusions of Chapter 5

The one way to increase performance of CELP based voice codecs is to adapt codec's codebooks for speech samples of specific language. Experiments performed on Lithuanian and Arabic languages show that:

- Codebook adaptation for Speex voice codec in average improves PESQ MOS scores about 0.2 points together improving performance stability for individual speech samples.
- The increase in Speex codec's performance is more noticeable at lower bitrates, by improving PESQ MOS values by about 0.35 MOS points.

General Conclusions

1. Performed investigations on performance of voice codecs show that the quality of Lithuanian and Arabic speech coding using CELP based voice codecs depends on specific language.
2. AMR and Speex codecs quality show bias towards English language, by exhibiting about 0.2 points higher average PESQ MOS values and about 0.05 lower spectral distortions.
3. AMR and Speex codecs subjective quality MOS is on average 0.22 lower for Lithuanian and 0.18 for Arabic languages in comparison with English language.
4. Fixed (innovation) codebook adaptation for Lithuanian and Arabic languages improves Speex codecs quality while coding Lithuanian and Arabic speech on average about 0.2 points together reducing its deviation about 0.1 points or 7–10 %.

References

3GPP TR 26.975 V5.0.0. 2002. Performance Characterization of the AMR Speech Codec (Release 5).

3GPP TS 26.071: Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Mandatory speech CODEC speech processing functions; AMR speech Codec; General description.

3GPP TS 26.073: "Adaptive Multi-Rate (AMR); ANSI C source code".

Andersen, S.; Kleijn, W.; Hagen, R; Linden, J.; Murthi, M.; Skoglund, J. 2002. iLBC-a linear predictive coder with robustness to packet losses. *Proc. IEEE Speech Coding Workshop*: 23–25.

ANSI S3.2-1989 (R1999) (1898). Method for Measuring the Intelligibility of Speech over Communications System.

Anskaitis, A. 2009. Koduoto balso kokybės tyrimas [Analysis of quality of coded voice signals]: *Doctoral Dissertation*. Vilnius Gediminas Technical University. Vilnius: Technika.

Barnett, P. W. 1999. Overview of speech intelligibility. *In Proceedings of the Institute of Acoustics*: 21.

Burnett, I. S. 1993. Hybrid Technology for Speech Coding. PhD. thesis; Chapter 3, University of Bath.

Burnett, S.; Parry, J. J. 1996. On the effects of accent and languages on low rate speech coders. *In Proceedings of Fourth International Conference on Spoken Language Processing*, ICSLP 96: 1: 291–294.

Campbell, J.P *et al*. 1990. The Proposed Federal Standard 1016 4,800 bps Voice Coder: CELP. *Speech Technology*: 58 – 64.

Cavander, A. *et al.* 2008. MobileASL: Intelligibility of sign language video over mobile phones, *Disability and Rehabilitation: Assistive Technology*: 3(1): 93–105.

Chen, J. H. 1995. Toll-quality 16 kb/s celp speech coding with very low complexity. *In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*: 9–12.

Cutler, A.; Carter, D. M. 1987. The predominance of strong initial syllables in English vocabulary. *Computer Speech and Language*: (2): 133–142.

Davidson, G.; Jamieson, D. G.; Parsa, V. and Price, M. C. 2002. Interaction of Speech Coders and Atypical Speech I: Effects on Speech Intelligibility. *J. Speech, Language, Hearing*: 45: 482–493.

Deler, J. R.; Hansen, J. H. L.; Proakis. J. P. 2000. Discrete-Time Processing of Speech Signals. *IEEE Press*.

Delogu, C.; Paoloni, A.; Pocci, P. 1991. New directions in the evaluation of voice Input-output systems. *Selected Areas in Communications*: 9(4): 566–573.

Donald, G.; Gersho, A. 1986. Complexity Reduction Methods for Vector Excitation Coding. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Donald, G.; Jamieson, D. G.; Parsa, V.; Price, M. C.; Till, J. 2002 Interaction of speech coders and atypical speech I: Effects on speech intelligibility. *Journal of Speech, Language and Hearing Research*: 45: 482–493.

Filipovič M. 2005. Atskirai tariamų lietuvių šnekos žodžių atpažinimo, grindžiamo dirbtiniais neuroniniais tinklais ir paslėptais markovo modeliais, tyrimai. PhD thesis, Vytauto Didžiojo universitetas.

Itani, M.; Paulikas. S. 2008. Influence of languages on CELP codecs performance. *Information Technology and Control*: 37(2): 141–144.

ITU-T Recommendaion. 1996. Subjective performance assessment of telephone-band and wideband digital codecs P.830

ITU-T Recommendation G.711: "Pulse code modulation (PCM) of voice frequencies".

ITU-T Recommendation H.324: "Terminal for low bit-rate multimedia communication".

ITU-T Recommendation P.50 Appendix I. 1998. Telephone transmission quality, telephone installations, local line networks. Objective measuring apparatus: Test signals.

ITU-T Recommendation P.563. 2004. Single ended method for objective speech quality assessment in narrowband telephony applications.

ITU-T. 1998. Telephone transmission quality, telephone installations, local line networks. Objective measuring apparatus: Test signals. ITU-T Recommendation P.830. P.50 Appendix I.

ITU-T. 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862.

ITU-T. 2004. Single ended method for objective speech quality assessment in narrowband telephony applications. ITU-T Recommendation P.563.

Joskowicz, J. *et al.* 2009. A Mathematical Model for Evaluating the Perceptual Quality of Video, *Lecture Notes in Computer Science* 5630: 1883–1893.

Kajackas, A.; Anskaitis, A. 2009. An Investigation of the Perceptual Value of Voice Frames, *Informatica*: 20(4): 487–498.

Kajackas, A.; Anskaitis, A.; Guršnys, D. 2009. Estimating Individual QoS, Traffic and QoS Management in Wireless Multimedia Networks. *COST290 Final Report: Lecture Notes in Electrical Engineering*: 31: 180–183.

Kleijn, W.B.; Haagen, J. 1994. Transformation and decomposition of speech signals for coding. *IEEE Sig. Proc. Letters*: 1(9): 136-138.

Kroon, P.; Atal, B. 1990 .Pitch predictors with high temporal resolution. *In Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*: 661–664.

Kurasova, O.; Dzemyda, G.; Vainoras, A. 2007. Parameter system for human physiological data representation and analysis. Pattern Recognition and Image Analysis. *Proceedings Third Iberian Conference*, IbPRIA 2007: 1: 209–216.

Laurinčiukaitė S. 2008. Lietuvių šnekos atpažinimo akustinis modeliavimas. PhD thesis, Vilniaus Gedimino technikos universitetas, [žiūrėta 2008 m. gruodžio 12 d.]. Prieiga per internetą: http://www.mii.lt/files/mii_dis_08_laurinciukaite.pdf.

Laurinčiukaitė, S.; Šilingas, D., Skripkauskas, M., & Telksnys, L. 2006. Lithuanian continuous speech corpus LRN 0.1: Design and potential applications. *Information Technology and Control:* (4), 431–440.

Maskeliunas, R.; Rudzionis, A.; Rudzionis, V. 2009. Analysis of the possibilities to adapt the foreign language speech recognition engines for the Lithuanian spoken commands recognition. Cross-Model Analysis of Speech, Gestures, Gaze and Facial Expressions. *COST Action 2102 International Conference. Revised Selected and Invited Papers*: 409-22|xx+434.

Paulikas, Š; Navakauskas, D. 2005. Restoration of voiced speech signals preserving prosodic features. *Speech Communication* 47(4): 457-468.

Simutis, R., Dilijonas, D., Bastina, L. 2009. Identification of unexpected behavior of an automatic teller machine using principal component analysis models. Business Information Systems Workshops 2009 : BIS 2009 international workshops, Poznan, Poland, April 2009 : revised papers. Book Series: *Lecture Notes in Business Information Processing*: 37: 53-61

Structured Codebook Design in CELP. *In Proceedings of International Mobile Satellite Conference*: 667–672.

Linde, Y.; Buzo, A.; Gray, R. M. 1980. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*: 28(1): 84–95.

Machoul, J.; Roucos, S. and Gish, H. 1985. Vector Quantization in Speech Coding. *In Proceedings of IEEE*: 73: 1551–1588.

Ojala, P.; Toukomaa, H.; Moriya, T.; Kunz, O. 1998. Report on the MPEG-4 speech codec verification tests. *Technical report*, MPEG Audio and Test subgroups.

Parry, J. J. 1995. Accent Classification for Speech Coding, Honours thesis, The University of Wollongong.

Parrya, J. J.; Burnett, I. S.; Chicharoa, J. F. 2000. Language-Specific Phonetic Structure and the Quantization of the Spectral Envelope of Speech. *Speech Communication*: 32(4): 229–250.

Parrya, J. J.; Burnett, I. S. 1996. On the effects of accent and language on low rate speech coders. *In Proceedings of Fourth International Conference on Spoken Language Processing, ICSLP 96:* 1: 291–294.

Paulikas, S.; Karpavičius, R. 2007. Application of linear prediction coefficients interpolation in speech signal coding. *Electronics and Electrical Engineering*, Kaunas, Technologija,: 80 (8): 39–42.

Raškinis, A., Raškinis, G.; Kapočiūtė, J. 2008. Diktoriaus skaitomas rišlaus teksto garsynas. *Technical report*, Vytauto Didžiojo Universitetas, [Referenced 12 of November 2009]. http://internet.ktu.lt/lt/mokslas/konf05/konf_- 02/IT2005/Sekc05.pdf.

Raškinis, A.; Raškinis, G.; Kazlauskienė, A. 2004. Lietuvių bendrinės šnekamosios kalbos garsų fondo kūrimo principai. *Lituanistica*: 60(4): 53–62.

Raškinis, A.; Raškinis, G.; Kazlauskienė, A. 2004. VDU bendrinės lietuvių šnekos universalus anotuotas garsynas. Technical report, Vytauto Didžiojo Universitetas, [žiūrėta 2009 m. lapkričio 12 d.]. Prieiga per internetą: http://elibrary.lt/resursai/Konferencijos/KTU_01/IT_2003/Sekcija09.pdf.

Rudžionis, A. 2001. Balso technologijų pasiekimai pasaulyje. [žiūrėta 2009 m. lapkričio 12 d.]. Prieiga per internetą: http://www.likit.lt/all/balso_tech/01_ivadas.htm: Informacinės visuomenės plėtros komitetas.

Schroeder, M.; Atal, B. 1984. Code-excited linear prediction(celp): High-quality speech at very low bit rates. I*n Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*: 937–940.

Schroeder, M.; Atal. B. 1985. Code-excited linear prediction(CELP):High-quality speech at very low bit rates. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP'85: 10: 937–472.

Soo J. S.; Pang, K. 1990. Multidelay block frequency domain adaptive filter: *IEEE Trans. on Acoustics, Speech and Signal Processing*, 38(2): 373–376.

Syrdal, A.K.; Bennett, W.R.; Greenspan, L. S. 1995. Applied Speech Technology. CRC Press.

Tiago, H. F.; Wai-Yip, C. 2009. Performance Study of Objective Speech Quality Measurement for Modern Wireless-VoIP Communications. *EURASIP Journal on Audio, Speech, and Music Processing*: ID 104382: 11.

Uvliden, A.; Bruhn, S.; Hagen, R. 1998. Adaptive multi-rate: A speech codec adapted to cellular radio network quality. *In Proceedings of 32nd ASILOMAR Conference*: 1: 343–347.

Valin J. M. 2007. The Speex Codec Manual (Version 1.2 Beta 3). Xiph.org Foundation.

Valin J. M.; Lefebvre, R. 2000. Bandwidth extension of narrowband speech for low bit-rate wideband coding. *In Proceedings IEEE Speech Coding Workshop*: 130–132.

Valin, J. M. 2004. The Speex Codec Manual (version 1.0.4).

Vindašius, A. 2010. Analysis of Quality of Service in Heterogenous Wireless Networks*: Doctoral Dissertation*. Vilnius Gediminas Technical University. Vilnius: Technika.

Žilinskienė, V. 2002. Statistical analysis of the morphology of Lithuanian administrative and publicistic styles. *Lituanistica*: 49(1): 106–116.

Žilinskienė,V. 2003. The use of grammatical forms in Lithuanian works of fiction. *Lituanistica*: 55(3): 75–84.

List of Publications by the Author on the Topic of the Dissertation

Papers in the Reviewed Scientific Journals

Paulikas, Š.; Itani, M. 2010. Lithuanian Speech Records Database for Voice Codecs Quality Assessment. *Information Technology and Control* 39(1): 38–42. ISSN 1392–124X. (Thomson ISI Web of Science).

Itani, M.; Paulikas, Š. 2008. Influence of languages on CELP codecs performance. *Information Technology and Control* 37(2):141–144. ISSN 1392–124X. (Thomson ISI Web of Science).

Papers in the Reviewed Scientific Publications

Paulikas, Š.; Itani, M. 2009. Assessment of CELP Codecs Quality in Multi-Lingual Environment. *The 9th IEEE International Symposium on Signal Processing and Information Technology*. December 14–17, 2009 – Ajman, UAE. 55–60. ISBN 978-1-4244-5950-6. (IEEExplore, Inspec).

Annex. Matlab Source Code for Codecs Performance Evaluation

```
Performance of AMR codec
function process_amr();
%process all data
%line=['pesq +8000 ' wavFileName '.wav ' wavFileName '_'
int2str(order) '_0' int2str(ii-B1) '_0' int2str(i)
'.wav\n'];

bps=[4750 5150 5900 6700 7400 7950 10200 12200];

Dir='./AMERICAN_ENGLISH/'
fnApp='A_eng_';
rezult=processdir(Dir,fnApp);
pesq_amr_us=rezult

Dir='./BRITISH_ENGLISH/'
fnApp='B_eng_';
rezult=processdir(Dir,fnApp);
pesq_amr_gb=rezult

Dir='./LITHUANIAN/'
```

```
fnApp='LTU_';
rezult=processdir(Dir,fnApp);
pesq_amr_ltu=rezult

Dir='./LITHUANIAN2/'
fnApp='Lt_';
rezult=processdir(Dir,fnApp);
pesq_amr_lt=rezult

Dir='./ARABIC/'
fnApp='Ar_';
rezult=processdir(Dir,fnApp);
pesq_amr_ar=rezult

save amr_all;
save  pesq_amr bps pesq_amr_lt pesq_amr_ltu pesq_amr_gb
pesq_amr_us pesq_amr_ar;



function rezult=processdir(Dir,fnApp);
mode=['MR475'; 'MR515'; ' MR59'; ' MR67'; ' MR74'; 'MR795';
'MR102'; 'MR122'];
rezult=zeros(16,8);
for dtx=1:8,
    ii=1;
    for i=1:8,
        wavFileName=[Dir fnApp 'f' int2str(i) '_8.wav'];
        %dos(['enc -dtx ' int2str(dtx) ' ' wavFileName '
tmp.amr']);
        %dos('dec tmp.amr tmp.wav');
        %Encoding:
        %sox test.wav -r 8000 test.sw
        %amrnb-encoder test.sw test.amr
        [ss rr]=dos(['/opt/local/bin/sox ' wavFileName ' -r
8000 ./tmp.sw']);
        [ss rr]=dos(['./amrenc -dtx ' mode(dtx,:) ' '
'./tmp.sw' ' ./tmp.amr']);
        %Decoding:
        %amrnb-decoder test.amr test.sw
        %sox -r 8000 test.sw test.wav
        [ss rr]=dos('./amrdec ./tmp.amr ./tmp.sw');
        [ss rr]=dos('/opt/local/bin/sox -r 8000 -c 1 -b
16 ./tmp.sw ./tmp.wav');
        [s r]=dos(['./pesq_mac +8000 ' wavFileName
' ./tmp.wav']);
```

```
        rezult(ii,dtx)=str2double(r);
        ii=ii+1;
    end
    for i=1:8,
        wavFileName=[Dir fnApp 'm' int2str(i) '_8.wav'];
        %dos(['enc -dtx ' int2str(dtx) ' ' wavFileName '
tmp.amr']);
        %dos('dec tmp.amr tmp.wav');
        %Encoding:
        %sox test.wav -r 8000 test.sw
        %amrnb-encoder test.sw test.amr
        [ss rr]=dos(['/opt/local/bin/sox ' wavFileName ' -r
8000 ./tmp.sw']);
        [ss rr]=dos(['./amrenc -dtx ' mode(dtx,:) ' '
'./tmp.sw' ' ./tmp.amr']);
        %Decoding:
        %amrnb-decoder test.amr test.sw
        %sox -r 8000 test.sw test.wav
        [ss rr]=dos('./amrdec ./tmp.amr ./tmp.sw');
        [ss rr]=dos('/opt/local/bin/sox -r 8000 -c 1 -b
16 ./tmp.sw ./tmp.wav');
        [s r]=dos(['./pesq_mac +8000 ' wavFileName
' ./tmp.wav']);
        rezult(ii,dtx)=str2double(r);
        ii=ii+1;
    end
end
```

```
Performance of Speex codec
function process_speex();
%process all data with speex

%speexenc --bitrate 4750 LTU_f1.wav LTU_f1.spx
%speexenc --bitrate 4750 LTU_m1_1.wav LTU_m1_1.spx

bps=[4750 5150 5900 6700 7400 7950 10200 12200];

Dir='./AMERICAN_ENGLISH/'
fnApp='A_eng_';
rezult=processdir(Dir,fnApp,bps);
pesq_speex_us=rezult;

Dir='./BRITISH_ENGLISH/'
fnApp='B_eng_';
rezult=processdir(Dir,fnApp,bps);
pesq_speex_gb=rezult;

Dir='./LITHUANIAN/'
fnApp='LTU_';
rezult=processdir(Dir,fnApp,bps);
pesq_speex_ltu=rezult;

Dir='./LITHUANIAN2/'
fnApp='Lt_';
rezult=processdir(Dir,fnApp,bps);
pesq_speex_lt=rezult;

Dir='./ARABIC/'
fnApp='Ar_';
rezult=processdir(Dir,fnApp,bps);
pesq_speex_ar=rezult

save speex_all
save pesq_speex bps pesq_speex_lt pesq_speex_ltu
pesq_speex_gb pesq_speex_us pesq_speex_ar;



function rezult=processdir(Dir,fnApp,bps);
rezult=zeros(16,8);
for dtx=1:8,
    ii=1;
    for i=1:8,
        wavFileName=[Dir fnApp 'f' int2str(i) '_8.wav'];
```

```
        [ss rr]=dos(['/usr/local/bin/speexenc --abr '
int2str(bps(dtx)) ' ' wavFileName ' ./tmp.spx']);
        [ss rr]=dos('/usr/local/bin/speexdec tmp.spx
tmp.wav');
        [s r]=dos(['./pesq_mac +8000 ' wavFileName
' ./tmp.wav']);
        rezult(ii,dtx)=str2double(r);
        ii=ii+1;
    end
    for i=1:8,
        wavFileName=[Dir fnApp 'm' int2str(i) '_8.wav'];
        [ss rr]=dos(['/usr/local/bin/speexenc --abr '
int2str(bps(dtx)) ' ' wavFileName ' ./tmp.spx']);
        [ss rr]=dos('/usr/local/bin/speexdec
tmp.spx ./tmp.wav');
        [s r]=dos(['./pesq_mac +8000 ' wavFileName
' ./tmp.wav']);
        rezult(ii,dtx)=str2double(r);
        ii=ii+1;
    end
end
```

```
3SQM scores
function process_3sqm();
%process all data with speex

%speexenc --bitrate 4750 LTU_f1.wav LTU_f1.spx
%speexenc --bitrate 4750 LTU_m1_1.wav LTU_m1_1.spx

bps=[4750 5150 5900 6700 7400 7950 10200 12200];

Dir='./AMERICAN_ENGLISH/'
fnApp='A_eng_';
rezult=processdir(Dir,fnApp,bps);
s3sqm_us=rezult;


Dir='./BRITISH_ENGLISH/'
fnApp='B_eng_';
rezult=processdir(Dir,fnApp,bps);
s3sqm_gb=rezult;

Dir='./ARABIC/'
fnApp='Ar_';
rezult=processdir(Dir,fnApp,bps);
s3sqm_ar=rezult;

Dir='./LITHUANIAN/'
fnApp='LTU_';
rezult=processdir(Dir,fnApp,bps);
s3sqm_ltu=rezult;

Dir='./LITHUANIAN2/'
fnApp='Lt_';
rezult=processdir(Dir,fnApp,bps);
s3sqm_lt=rezult;

save s3sqm_all
save  s3sqm s3sqm_lt s3sqm_ltu s3sqm_gb s3sqm_us s3sqm_ar;



function rezult=processdir(Dir,fnApp,bps);
rezult=zeros(16,1);
    ii=1;
    for i=1:8,
        wavFileName=[Dir fnApp 'f' int2str(i) '_8.wav'];
        [s r]=dos(['./p563 ' wavFileName]);
```

```
        rezult(ii)=str2double(r);
        ii=ii+1;
    end
    for i=1:8,
        wavFileName=[Dir fnApp 'm' int2str(i) '_8.wav'];
        [s r]=dos(['./p563 ' wavFileName]);
        rezult(ii)=str2double(r);
        ii=ii+1;
    end
```

```
Spectral distortions for AMR
function process_mse_amr();
%process all data
%line=['pesq +8000 ' wavFileName '.wav ' wavFileName '_'
int2str(order) '_0' int2str(ii-B1) '_0' int2str(i)
'.wav\n'];

bps=[4750 5150 5900 6700 7400 7950 10200 12200];

Dir='./ARABIC/'
fnApp='Ar_';
[snr,d]=processdir(Dir,fnApp);
snr_amr_ar=snr;
d_amr_ar=d;
%load mse_amr_all
%save mse_amr_all;
%save  mse_amr bps mse_amr_lt mse_amr_ltu mse_amr_gb
mse_amr_us mse_amr_ar;
%return

Dir='./AMERICAN_ENGLISH/'
fnApp='A_eng_';
[snr,d]=processdir(Dir,fnApp);
snr_amr_us=snr;
d_amr_us=d;

Dir='./BRITISH_ENGLISH/'
fnApp='B_eng_';
[snr,d]=processdir(Dir,fnApp);
snr_amr_gb=snr;
d_amr_gb=d;

Dir='./LITHUANIAN/'
fnApp='LTU_';
[snr,d]=processdir(Dir,fnApp);
snr_amr_lt=snr;
d_amr_lt=d;

% Dir='./LITHUANIAN2/';
% fnApp='Lt_';
% rezult=processdir(Dir,fnApp);
% mse_amr_lt=rezult


save snrD_amr_all;
```

```
save  snrD_amr bps snr_amr_lt snr_amr_gb snr_amr_us
snr_amr_ar d_amr_lt d_amr_gb d_amr_us d_amr_ar;



function [SNR, D]=processdir(Dir,fnApp);
mode=['MR475'; 'MR515'; ' MR59'; ' MR67'; ' MR74'; 'MR795';
'MR102'; 'MR122'];
SNR=zeros(16,8);
D=zeros(16,8);
for dtx=1:8,
    ii=1;
    for i=1:8,
        wavFileName=[Dir fnApp 'f' int2str(i) '_8.wav'];
        %Encoding:
        %sox test.wav -r 8000 test.sw
        %amrnb-encoder test.sw test.amr
        [ss,rr]=dos(['/opt/local/bin/sox ' wavFileName ' -r
8000 ./tmp.sw']);
        [ss,rr]=dos(['./amrenc -dtx ' mode(dtx,:) ' '
'./tmp.sw' ' ./tmp.amr']);
        %Decoding:
        %amrnb-decoder test.amr test.sw
        %sox -r 8000 test.sw test.wav
        [ss,rr]=dos('./amrdec ./tmp.amr ./tmp.sw');
        [ss,rr]=dos('/opt/local/bin/sox -r 8000 -c 1 -b
16 ./tmp.sw ./tmp.wav');
        %[s r]=dos(['pesqmain.exe +8000 ' wavFileName '
tmp.wav']);
        %rezult(ii,dtx+1)=str2double(r);

        s=wavread(wavFileName);
        s_tmp=wavread('./tmp.wav');

        len=min([length(s) length(s_tmp)]);
%         plot([s(1:len) s(1:len)-s_tmp(1:len)])
%         sum((s(1:len)-s_tmp(1:len)).^2)
%         sum(s(1:len).^2)

        snr=10*log10(sum(s(1:len).^2)./sum((s(1:len)-
s_tmp(1:len)).^2));

d=sqrt(1/len*sum(10*log10(fft(s(1:len)).^2./fft(s_tmp(1:len
)).^2)));
        SNR(ii,dtx)=snr;
        D(ii,dtx)=real(d);
```

```
        ii=ii+1;
    end
    for i=1:8,
        wavFileName=[Dir fnApp 'm' int2str(i) '_8.wav'];
        %Encoding:
        %sox test.wav -r 8000 test.sw
        %amrnb-encoder test.sw test.amr
        [ss,rr]=dos(['/opt/local/bin/sox ' wavFileName ' -r
8000 ./tmp.sw']);
        [ss,rr]=dos(['./amrenc -dtx ' mode(dtx,:) ' '
'./tmp.sw' ' ./tmp.amr']);
        %Decoding:
        %amrnb-decoder test.amr test.sw
        %sox -r 8000 test.sw test.wav
        [ss,rr]=dos('./amrdec ./tmp.amr ./tmp.sw');
        [ss,rr]=dos('/opt/local/bin/sox -r 8000 -c 1 -b
16 ./tmp.sw ./tmp.wav');

        s=wavread(wavFileName);
        s_tmp=wavread('./tmp.wav');
        len=min([length(s) length(s_tmp)]);
        snr=10*log10(sum(s(1:len).^2)./sum((s(1:len)-
s_tmp(1:len)).^2));

d=sqrt(1/len*sum(10*log10(fft(s(1:len)).^2./fft(s_tmp(1:len
)).^2)));
        SNR(ii,dtx)=snr;
        D(ii,dtx)=real(d);
        ii=ii+1;
    end
end
```

```
Spectral distortions for Speex
function process_mse_amr();
%process all data
%line=['pesq +8000 ' wavFileName '.wav ' wavFileName '_'
int2str(order) '_0' int2str(ii-B1) '_0' int2str(i)
'.wav\n'];

bps=[4750 5150 5900 6700 7400 7950 10200 12200];

Dir='./ARABIC/'
fnApp='Ar_';
[snr,d]=processdir(Dir,fnApp,bps);
snr_speex_ar=snr;
d_speex_ar=d;
%load mse_speex_all
%save mse_speex_all;
%save  mse_speex bps mse_speex_lt mse_speex_ltu
mse_speex_gb mse_speex_us mse_speex_ar;
%return

Dir='./AMERICAN_ENGLISH/'
fnApp='A_eng_';
[snr,d]=processdir(Dir,fnApp,bps);
snr_speex_us=snr;
d_speex_us=d;

Dir='./BRITISH_ENGLISH/'
fnApp='B_eng_';
[snr,d]=processdir(Dir,fnApp,bps);
snr_speex_gb=snr;
d_speex_gb=d;

Dir='./LITHUANIAN/'
fnApp='LTU_';
[snr,d]=processdir(Dir,fnApp,bps);
snr_speex_lt=snr;
d_speex_lt=d;

% Dir='./LITHUANIAN2/';
% fnApp='Lt_';
% rezult=processdir(Dir,fnApp);
% mse_speex_lt=rezult


save snrD_speex_all;
```

```
save  snrD_speex bps snr_speex_lt snr_speex_gb snr_speex_us
snr_speex_ar d_speex_lt d_speex_gb d_speex_us d_speex_ar;



function [SNR, D]=processdir(Dir,fnApp,bps);
mode=['MR475'; 'MR515'; ' MR59'; ' MR67'; ' MR74'; 'MR795';
'MR102'; 'MR122'];
SNR=zeros(16,8);
D=zeros(16,8);
for dtx=1:8,
    ii=1;
    for i=1:8,
        wavFileName=[Dir fnApp 'f' int2str(i) '_8.wav'];

        [ss rr]=dos(['/usr/local/bin/speexenc --abr '
int2str(bps(dtx)) ' ' wavFileName ' ./tmp.spx']);
        [ss rr]=dos('/usr/local/bin/speexdec tmp.spx
tmp.wav');
        s=wavread(wavFileName);
        s_tmp=wavread('./tmp.wav');

        len=min([length(s) length(s_tmp)]);
%         plot([s(1:len) s(1:len)-s_tmp(1:len)])
%         sum((s(1:len)-s_tmp(1:len)).^2)
%         sum(s(1:len).^2)

        snr=10*log10(sum(s(1:len).^2)./sum((s(1:len)-
s_tmp(1:len)).^2));

d=sqrt(1/len*sum(10*log10(fft(s(1:len)).^2./fft(s_tmp(1:len
)).^2)));
        SNR(ii,dtx)=snr;
        D(ii,dtx)=real(d);
        ii=ii+1;
    end
    for i=1:8,
        wavFileName=[Dir fnApp 'm' int2str(i) '_8.wav'];

        [ss rr]=dos(['/usr/local/bin/speexenc --abr '
int2str(bps(dtx)) ' ' wavFileName ' ./tmp.spx']);
        [ss rr]=dos('/usr/local/bin/speexdec tmp.spx
tmp.wav');
        s=wavread(wavFileName);
        s_tmp=wavread('./tmp.wav');
        len=min([length(s) length(s_tmp)]);
```

```
        snr=10*log10(sum(s(1:len).^2)./sum((s(1:len)-
s_tmp(1:len)).^2));

d=sqrt(1/len*sum(10*log10(fft(s(1:len)).^2./fft(s_tmp(1:len
)).^2)));
        SNR(ii,dtx)=snr;
        D(ii,dtx)=real(d);
        ii=ii+1;
    end
end
```

Mohamad ITANI

ANALYSIS AND IMPROVEMENT
OF MULTILINGUAL SPEECH CODING

Doctoral Dissertation

Technological Sciences,
Electrical and Electronics Engineering (01T)

ĮVAIRIŲ KALBŲ KALBOS SIGNALŲ
KODAVIMO ANALIZĖ IR GERINIMAS

Daktaro disertacija

Technologijos mokslai,
elektros ir elektronikos inžinerija (01T)

2011 02 07. 8,5 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“,
Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Ciklonas“
J. Jasinskio g. 15, 01111 Vilnius.