Computers and Electrical Engineering 49 (2016) 9–24

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

A two-level multi-gene genetic programming model for speech quality prediction in Voice over Internet Protocol systems[☆]

Farhad Rahdari[a,*], Mahdi Eftekhari[b], Reza Mousavi[c]

[a] Department of Computer and IT, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran
[b] Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
[c] Department of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran

ARTICLE INFO

Article history:
Received 19 January 2015
Revised 25 October 2015
Accepted 26 October 2015
Available online 19 November 2015

Keywords:
VoIP
2-level structure
Speech quality prediction
Non-intrusive
Multi-gene GP
PESQ

ABSTRACT

The main aim of this study is to develop a low-complexity non-intrusive quality prediction model in Voice over Internet Protocol (VoIP) systems. In order to gain this goal, a 2-level structure for predicting the quality of speech is proposed. Furthermore, the capabilities of multi-gene genetic programming are investigated through developing a number of parallel models and different feature vectors. These models are utilized in two hierarchical levels to construct the final model. To consider the transmission media and speech signal characteristics in quality measurement process, both network impairments and per-frame features are employed simultaneously for developing models. Several experiments are performed based on the proposed structure while different combinations of speech feature types in the cases of noise free and noisy speech signals are examined. The obtained results indicate that using parallel models in a 2-level structure enhances the accuracy of derived models as compared with 1-level structure and common single-gene GP models.

1. Introduction

In the world of real-time technologies, the perceived quality plays a critical role in increasing user satisfaction level. Hence, developing non-intrusive models to predict speech quality has been an important research subject during the past few years. The non-intrusive term refers to those methods do not need the reference signal in the process of estimating the quality. Originally, speech quality can be measured by performing reliable subjective tests. These methods utilize Mean Opinion Score (MOS) to categorize speech quality from poor to excellent based on human listener opinions [1]. Nevertheless, performing costly and time-consuming tests makes these methods unsuitable for real-time applications such as Voice over Internet Protocol (VoIP). Therefore, the objective models have been proposed to predict test results (i.e. MOS) from relevant parameters instead of human judgment. This type of evaluation methods can be classified as intrusive or non-intrusive [2]. The intrusive methods estimate the quality from calculation of the difference between clean and output degraded signals. The difference value (i.e. distortion) is then mapped into MOS value to obtain the predicted quality [2]. The most common intrusive method which formed as ITU-T P.862, is Perceptual Evaluation of Speech Quality (PESQ) [3]. The intrusive methods provide good correlation with subjective results.

[☆] Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Jia Hu.
[*] Corresponding author. Tel.: +98 3432228468.
E-mail addresses: rahdarifar@icst.ac.ir, rahdarifar@gmail.com (F. Rahdari), m.eftekhari@mail.uk.ac.ir (M. Eftekhari), r.mousavi@student.kgut.ac.ir (R. Mousavi).

http://dx.doi.org/10.1016/j.compeleceng.2015.10.008
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

However, the major drawback is the need for original signal which make them unusable for live traffics. In order to overcome this weakness, non-intrusive methods have been proposed which measure the quality from speech features or transmission parameters. The signal-based methods measure the quality directly from analysis of the degraded version of received speech. For example, the ITU-T P.563 [4] utilizes various speech features to predict overall quality according to three main distortion classes. Parameters-based methods, in other hands, calculate the speech quality by using a number of parameters relevant to communication network. For instance, the E-model [5] uses some transmission impairments to develop a mathematical linear relation for predicting speech quality score. This model provides a quality rating factor *R* on a scale of 0–100.

In recent years, more attention has been paid to the use of Artificial Intelligence (AI) algorithms for developing non-intrusive quality assessment models. In [6] a parametric-based model by using Artificial Neural Network (ANN) for speech quality estimation was proposed. Falk et al. develop a speech quality prediction model using a set of extracted speech features together with Gaussian Mixture Models (GMMs) [7]. An auditory model for non-intrusive speech quality estimation was proposed in [8], which is based on temporal envelope representation of speech signal. Researchers in [9] employed the per-frame features which commonly are used in speech coding to predict speech quality. The proposed model utilized the degree of consistency between speech features and GMM of original signal to estimate overall quality. Raja [10] explored the capabilities of the Genetic Programming (GP) by developing a symbolic form model based on common single-gene GP. A non-intrusive objective measurement for estimating the quality of speech is proposed based on fuzzy Gaussian Mixture Model (GMM) and Support Vector Regression (SVR) was proposed in [11] for both narrowband and wideband speech. In [12,13], different types of neural network were utilized to estimate speech quality, non-intrusively. These approaches employed the PESQ results to prepare dataset and predict overall MOS score. Researchers in [14] utilized Mel Filter Bank energies as the desired features and develop a non-intrusive prediction model based on Support Vector Regression (SVR). In [15] the capabilities of Bayesian learning method was studied for estimating the quality of VoIP by using several number of network impairments such as codec type, packet loss, gender and language of speaker. A parameter-based approach by using common single-gene GP method for quality estimation was introduced in [16]. A fuzzy-based model was also developed in [17] which uses a hybrid of GA and Neuro-Fuzzy for estimating quality of speech in IP networks. In [18] a modified version of E-model (ITU-G.107) was introduced by adding two new parameters related to delay impairment. In order to exploit the advantages of ensemble learning method, different base learners were learned for developing quality model in [19]. The proposed method uses the MFCC features which are extracted from different frequency sub-bands of original signal. Also, the Discrete Wavelet Transform (DWT) was utilized to decompose original signal into different sub-bands. In [20] different factors including culture, ageing, and language were used to measure perception of Quality of Service (QoS) of IP applications. The paper [21] exploited the subjective test results to develop a mathematical model for quality measurement based on native Thai users.

The aim of this study is to develop a low complexity model for predicting the quality in VoIP systems. In this way, similar to [10,16], the GP algorithm is used as one of the well-known AI algorithms for developing quality measurement model. The GP family algorithms are the biologically-inspired techniques which attempt to find a solution (program) for a problem in a symbolic form through a number of genetic operations. In contrast to mentioned studies, a unique type of GP algorithm which utilizes the multi-gene individuals instead of common single-gene ones is utilized in this study. Using multi-gene GP is due to its ability in developing more accurate and efficient symbolic regression models in comparison with the common GP. Furthermore, in the most studies in the field of quality measurement, the proposed models are parameters-aware or signal-aware, exclusively. Against these approaches, this paper proposes a signal-aware model which is not unaware of transmission media conditions. This hybrid method utilizes the transmission impairments and speech features simultaneously to derive desired quality model. Another aspect is the complexity of developed models which is a major problem in the most of studies due to large feature vectors. In order to overcome this challenge, a 2-level structure is proposed which utilizes the parallel models at the first level. The parallelism leads to dimension reduction which avoids the curse of dimensionality. Utilizing parallel models also provides the ability for exploring the effect of combining speech features in improving model performance. Based on proposed structure, the second level calculates the overall quality score by aggregating the intermediate quality scores that are obtained at parallel models. For this, a different number of the network impairments and speech features are organized in a single feature vector to derive quality measurement model. The efficiency of proposed method is measured by performing several numbers of experiments in the cases that original (noise-free) and noisy speech signals are used at the sender side of VoIP systems.

The rest of this paper is organized as follows: in Section 2, we describe two common types of objective speech quality measurement methods including intrusive and non-intrusive briefly. In Section 3, various speech signal features and transmission impairments are described. Section 4 introduces MGGP and then describes the proposed 2-level structure for developing MGGP-based prediction model. Section 5 includes simulation system setup and dataset preparation. In Section 6, different experiments are performed according to different evolved MGGP model and results are compared. Section 7 concludes the paper.

2. Objective speech quality assessment

The objective methods attempts to predict subjective scores through an objective process based on human perception. The objective methods can be categorized in three main types: (1) The time domain methods like signal-to-noise ratio (SNR) and segmental SNR (SSNR) [2] use the original and degraded signals difference in time domain to obtain quality score, (2) The spectral domain methods use the parameters of speech production model to estimate quality. The Itakura–Saito (IS) measure, Spectral Distortion (SD), Linear Predictive Coding (LPC) parameter distance are examples of these methods [2], and (3) The perceptual domain methods exploit the operation of human auditory perception system for quality measurement. As shown in Fig. 1,