Digital Signal Processing 70 (2017) 114–124

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Non-intrusive speech quality estimation as combination of estimates using multiple time-scale auditory features

Rajesh Kumar Dubey [a,*], Arun Kumar [b]

[a] Department of Electronics & Communication Engg., Jaypee Institute of Information Technology, Noida (UP)-201307, India
[b] Centre for Applied Research in Electronics, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi-110016, India

ARTICLE INFO

Article history:
Available online 4 August 2017

Keywords:
Non-intrusive
Speech quality
Multiple time-scale features
Auditory model
Degraded speech

ABSTRACT

The human auditory system is modeled by different auditory models representing the distribution of speech sound energy in different channels across the cochlea using filter-banks of different bandwidths. In previous algorithms of non-intrusive speech quality evaluation, auditory features are determined using these auditory models on per frame basis and then averaged over the entire speech utterance. In these approaches, the effect of impulsive noise and other non-stationary noise effects get averaged over the utterance. To include the variations in the features of speech over time in the speech utterance, a multiple time-scale features approach has been proposed as the speech features vary from frame to frame that accounts for variation of noise characteristics over the speech utterance and thus its affect on quality mapping. In this work, non-intrusive speech quality evaluation has been done as an optimal linear combination of quality mapping called objective mean opinion score (MOS), computed using multiple time-scale estimates of features. The objective MOS of each of the multiple time-scale estimates (the combination of multiple active speeches) are obtained using a probabilistic approach. The overall objective MOS of the speech utterance is computed by taking the optimal linear combination of the estimated objective MOS using multiple time-scale estimates of features, where the optimality is based on the minimum mean square error (MMSE) criterion or correlation maximization criterion. The results are given in terms of Pearson's correlation coefficient and root mean square error (RMSE) between the subjective MOS and the estimated overall objective MOS for three different standard databases. The results have been compared with a single time-scale features approach, the ITU-T Recommendation P.563 and recent algorithms.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

The performance evaluation of codecs in different speech processing algorithms, and monitoring and maintenance of the quality of service at different nodes in modern telecommunication networks require speech quality evaluation. One aspect of this requirement is to evaluate the speech quality objectively and continuously. The allocation of proper bandwidth or other remedial measures can be taken if the quality of speech falls below a desired level. The ideal method is through subjective listening tests according to absolute category rating (ACR) method called ITU-T Recommendation P.800 published in Aug. 1996 [1], where speech material is played and the mean value of the opinion scores of listeners is considered as the measure of quality, called the subjective mean opinion score (MOS). But this approach is not practical from system automation point of view as it is very time consuming and expensive. Thus, objective speech quality assessment methods are important in estimating speech quality. This measure of speech quality is called the objective MOS that should correlate well with the subjective MOS [2]. The objective speech quality measurements are done in two ways: intrusive or non-intrusive. In intrusive method (also called two-sided or double ended), the original clean speech utterance is required as a reference but in non-intrusive method (also called one-sided or single ended), only the received (or degraded) speech signal is used to estimate the quality, i.e. the objective MOS as shown in Fig. 1. Thus, a non-intrusive method is suitable for speech quality measurement at any node of the telecommunication network. Also, non-intrusive method of speech quality measurement is suitable for system automation and real-time applications where the original clean speech signal is practically impossible to obtain such as mobile communications, telephonic communications, direct-to-home (DTH) signal of television (TV), VoIP signal, etc. The percep-

* Corresponding author.
E-mail addresses: rajeshk_dubey@yahoo.com (R.K. Dubey), arunkm@care.iitd.ac.in (A. Kumar).
URL: http://care.iitd.ac.in/ (A. Kumar).

http://dx.doi.org/10.1016/j.dsp.2017.07.020
1051-2004/© 2017 Elsevier Inc. All rights reserved.

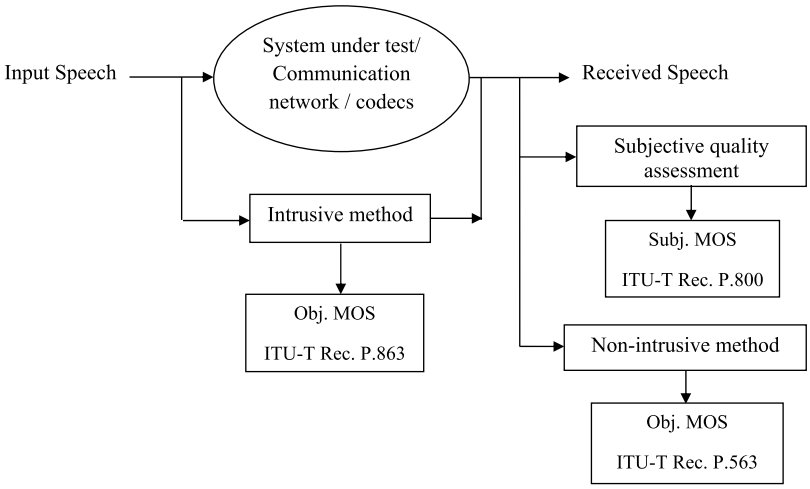

Fig. 1. Subjective and objective assessment of speech quality.

tual linear prediction (PLP) coefficients and perceptually weighted Bark spectrum which are speaker independent, are used for the first non-intrusive signal based model proposed in 1994, by Liang and Kubichek [3]. In this method, the artificial reference parameters corresponding to high speech quality are constructed from a clean source speech material and these parameters are compared with the degraded speech parameters. The visual features of the spectrogram of degraded speech are utilized in [4] by assuming that a "good" quality sentence will have discrete and dominant features while a "bad" quality sentence will typically have uniform distribution of energy in the spectrogram. In this method, average features of block-wise variance and dynamic range are calculated from the spectrogram of typically 10 to 30 ms and 50 to 500 Hz time-frequency resolution. The estimation of quality of network degraded speech by the use of vocal-tract modeling technique is given in [5]. In this method, the parameterized data are used to generate physiology based rules for error assessment utilizing cavity tracking techniques and context based error spotting. The ITU-T presented the Recommendation P.563 in 2004 for non-intrusive speech quality assessment [6], where speech quality assessment is done by modeling the vocal tract as a series of tubes with abnormal variations of the tube taken as degradation, re-constructing a clean intermediate reference signal from the distorted speech signal, and mapping of different distortion-specific parameters (noise, temporal clipping, robotization). The speech features are directly derived from speech coders without considering any degradation model and these features are mapped to the objective MOS using Gaussian mixture model (GMM) approach is described in [7]. The 11 per frame local features such as spectral flatness, spectral centroid, excitation variance, speech variance, pith-period, their time derivatives, and spectral dynamics are computed and statistical properties like mean, variance, skewness and kurtosis of these per frame features are used in this algorithm after dimensionality reduction to 14 features per speech utterance. The functional role of the human auditory system and the articulator system characteristics in the form of temporal envelope representation of speech have been utilized in the auditory non-intrusive quality estimation (ANIQUE) model [8]. The computational modeling of human auditory system, which takes into account filtering, detection and compression in the cochlea, has been done in [9]. In [10], non-intrusive objective speech quality evaluation has been done using GMM mapping of features of degraded speech. The features are obtained from Lyon's auditory model that takes into account the critical band phenomenon and the effect of masking of human auditory system. The GMM mapping of the combinations of several features such as Lyon's auditory features, mel-frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) features are investigated for non-intrusive speech quality evaluation in [11].

A limitation of current methods is that the features used for speech quality measurement are computed over an entire speech utterance and then mapped to an objective quality rating score. Let us call these methods as "single time-scale methods." The use of features over entire speech utterances in the single time-scale methods averages the effect of different types of degradations that are locally present at different locations within the speech utterance. In order to include the temporal masking phenomena of the human auditory system in the listening process and to cater for the effect of short-time transient additive noise/degradation present at different locations in the speech utterance in speech quality evaluation problem, a multiple time-scale features approach has been proposed [12]. The features are computed for multiple time-scale that capture both the local features at different time-scale within the speech utterance and the global features at a single time-scale corresponding to the entire speech utterance. It is hypothesized that the different characteristics of noise/degradation locally present at different locations in the speech signal, for example transients, but not in the entire speech utterance as a statistically stationary phenomenon, affects the perception of listeners and their quality ratings in a different way. If noise/degradation is predominantly present at the beginning of the speech utterance, then the listeners make their quality opinion in a different way as compared to the cases when noise/degradation is present in the middle or the end of the speech utterance or is uniformly spread over the entire speech utterance. Thus, the different multiple time-scale features over the utterance should have different contributions to the objective quality score mapping in a speech utterance. This hypothesis can be incorporated in an algorithm by computing the optimal weights of different multiple time-scale estimates and the overall objective MOS of a speech utterance is computed as a linear combination of the weighted objective quality score of the multiple time-scale estimates. In this work, non-intrusive objective speech quality assessment has been studied for narrowband telephonic speech using optimal linear combination of multiple time-scale estimates of Lyon's auditory features and other features such as MFCC and LSF by Gaussian Mixture Model (GMM) mapping. The active speech segments are obtained from the speech utterance by passing it through a voice activity detection (VAD) algorithm [13]. The multiple time-scale auditory features are computed using Lyon's auditory model for different combinations of active speech segments. The multiple time-scale active speech combinations are made by concatenating different contiguous active speech segments together. It is assumed that if

Download English Version:

https://daneshyari.com/en/article/4973766

Download Persian Version:

https://daneshyari.com/article/4973766

Daneshyari.com