Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 34–48

www.elsevier.com/locate/specom

Speech quality assessment using 2D neurogram orthogonal moments

Wissam A. Jassim[a,*], Muhammad S.A. Zilany[b]

[a] Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
[b] Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia

Received 10 June 2015; received in revised form 24 February 2016; accepted 28 March 2016
Available online 23 April 2016

Abstract

This study proposes a new objective speech quality measure using the responses of a physiologically-based computational model of auditory nerve (AN). The population response of the model AN fibers to a speech signal is represented by a 2D neurogram, and features of the neurogram are extracted by orthogonal moments. A special type of orthogonal moment, the orthogonal Tchebichef-Krawtchouk moment, is used in this study. The proposed measure is compared to the subjective scores from two standard databases, the NOIZEUS and the supplement 23 to the P series (P.Sup23) of ITU-T Recommendations. The NOIZEUS database is used in the assessment of 11 speech enhancement algorithms whereas the P.Sup23 database is used in the ITU-T 8 kbit/s codec (Recommendation G.729) characterization test. The performance of the proposed speech quality measure is also compared to the results from some traditional objective quality measures. In general, the proposed neural-response-based metric yielded better results than most of the traditional acoustic-property-based quality measures. The proposed metric can be applied to evaluate the performance of various speech-enhancement algorithms and compression systems.

Keywords: Speech quality assessment; Neurogram; Auditory-nerve model; Orthogonal moments; Discrete Tchebichef–Krawtchouk Transform DTKT; PESQ; POLQA.

1. Introduction

Speech is the major means of communication between people. In many situations, however, the speech signal is degraded, and only a limited transfer of information is obtained. This degradation may be due to factors related to the speaker, listener, and the type of speech, but in most situations, it is due to the transmission of the speech signal from the speaker to the listener (Steeneken, 1992). The distortion of speech could be measured in terms of different attributes, including quality and intelligibility. Quality measures assess "how" a speaker produces an utterance, and includes attributes such as natural, raspy, hoarse, and scratchy. Intelligibility measures assess "what" the speaker said, i.e., the meaning or the content of the spoken words (Loizou, 2011). Different assessment methods are used to evaluate quality and intelligibility of speech.

Assessment of speech quality can be done using subjective listening tests or objective measures. Subjective evaluation involves comparison of original and processed speech signals by a group of listeners who are asked to rate the quality of speech along a pre-determined scale. Objective evaluation may also involve a mathematical comparison of the original and processed speech signals (Loizou, 2013). The correlation with the scores from subjective listening tests determines the validity of any objective measure. The flow chart in Fig. 1 shows the general classification of speech-performance consisting of subjective and intrusive objective measures. Efforts have been made in the last few decades to develop objective speech quality and intelligibility measures. However, in this study we focus only on the intrusive (full-reference) speech quality metrics.

The frequency-weighted segmental signal-to-noise ratio (fwsegSNR) (Tribolet et al., 1978), audio quality assessment based on a model of auditory perception (PEMO-Q) (Huber and Kollmeier, 2006) and the perceptual evaluation of speech quality (PESQ) (ITU-T recommendation P.862, 2001) are well-known examples of objective speech quality measures.

* Corresponding author.
E-mail addresses: binaye2001@yahoo.com (W.A. Jassim), zilany@um.edu.my (M.S.A. Zilany).

http://dx.doi.org/10.1016/j.specom.2016.03.004
0167-6393/

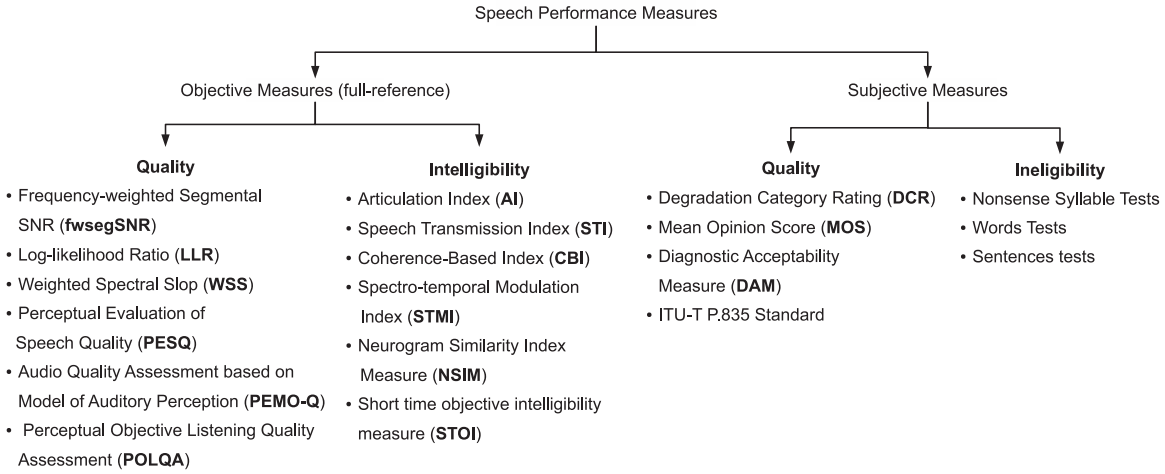

Fig. 1. Classification flowchart for the traditional measures of speech performance.

The mean opinion score (MOS) (Rothauser et al., 1969) and degradation category rating (DCR) (Panzer et al., 1993) are commonly-used subjective tests of speech quality. Most of the traditional objective quality measures use features of the original (clean) and distorted speech in the time or frequency domain to measure the similarity index. For example, the fwsegSNR measure is based on the geometric mean of the SNR across all frames of the speech signal (Loizou, 2013). Some objective measures were proposed based on the spectral distance of the linear predictive coding (LPC) models. The log-likelihood ratio (LLR) measure is an example of the method that uses all-pole models derived from LPC coefficients to predict the similarity between clean and enhanced speech signals (Quackenbush et al., 1988). The weighted spectral slope (WSS) measure is designed to determine the differences in the locations of formants or spectral peaks by finding the spectral slope of each band of signal frequencies (Klatt, 1982).

The PESQ measure is one of the most successful quality measures which is suitable for a wide range of distortions, packet loss, signal delays, and codec distortions (Loizou, 2013). It was recommended by the telecommunication standardization sector of the international telecommunication union (ITU) as a standard objective quality measure (P.862 standard). This method uses the time alignment technique in order to compensate for the delay between the original and distorted signals. PESQ employs a transformation which involves equalization for linear filtering and gain variation in the system to obtain the loudness spectra (ITU-T recommendation P.862, 2001; Rix et al., 2001).

Perceptual objective listening quality assessment (POLQA) is a new standard measure adopted by ITU-T as Recommendation P.863 in 2011 (Beerends et al., 2013; ITU-T, 2014). It addresses some of the known limitations of PESQ such as time alignment and warped speech for narrowband, wideband, and super-wideband speech (Hines et al., 2015). The main element of POLQA algorithm employed is the perceptual model which takes into account masking effects of the human hearing. The perceptual model is based on the idealization approach that removes low levels of noise in the reference input signal and optimizes the timbre. It also includes modeling the impact of playback levels on the perceived quality and a major split in the processing of low and high levels of distortion (Beerends et al., 2013).

Hu and Loizou (2008) assessed the performance of four acoustic-signal-property-based quality measures (SNR, LLR, WSS, and PESQ) to predict subjective scores for the quality of noisy speech enhanced by noise suppression algorithms such as spectral subtraction, subspace, statistical-model based, and Wiener algorithms. This study also proposed a composite objective measure by linearly combining the four individual objective measures. The composite measure was designed to predict the quality of three aspects of denoised speech: signal distortion, background noise, and overall quality. The subjective quality scores were previously reported in Hu and Loizou (2007) and Hu and Loizou (2006) (based on the methodology of the ITU-T P.835 recommendation).

Kates and Arehart proposed the hearing-aid speech quality index (HASQI) to predict the effects of nonlinear distortion, linear filtering, and noise on speech quality for both normal-hearing and hearing-impaired listeners (Kates and Arehart, 2010; Kressner et al., 2013). The proposed index is based on a cochlear model that incorporates elements of impaired hearing. It is composed of the product of two components. The first component is to predict quality with additive noise and nonlinear processing based on time-frequency cochlear representations using a basic cochlear model. The linear filtering and spectral changes, on the other hand, are captured by second component. The HASQI achieved a very high performance with training and testing datasets. The generalizability of HASQI was investigated in Kressner et al. (2013) by testing its ability to predict the subjective quality ratings of normal-hearing listeners of a dataset on which it was not trained. The reported results showed that the prediction performance of HASQI was comparable to the PESQ.

In this paper, a neural-response-based approach that uses the responses of the a nonlinear AN model to estimate the