Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 71 (2015) 1–9

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions

Peter Počta [a,*], John G. Beerends [b]

[a] Dept. of Telecommunications and Multimedia, FEE, University of Žilina, SK-01026 Žilina, Slovakia
[b] TNO, P.O. Box 96800, NL-2509 JE The Hague, The Netherlands

Received 27 January 2015; received in revised form 19 March 2015; accepted 1 April 2015
Available online 9 April 2015

Abstract

This paper investigates the impact of different telephone channels, represented by impairments as introduced by modern telecommunication networks (e.g. speech coding, bandwidth limitation, packet loss, etc.), on the intelligibility of synthesized speech. Both subjective and objective assessments are used. Two different speech intelligibility prediction models, namely PESQ Intelligibility and POLQA Intelligibility, are evaluated by comparing the predictions with subjectively obtained intelligibility scores. The results show that all the investigated degradations seriously impact the intelligibility of the synthesized speech measured subjectively. Furthermore it is shown that PESQ Intelligibility provides too low correlations between predicted objective measurements and subjective scores for accurate prediction of speech intelligibility while POLQA Intelligibility is capable of providing good intelligibility predictions in the case that a closed response experimental set up is used.

Keywords: Speech intelligibility; Synthesized speech; Telecom degradations; CVC test; POLQA Intelligibility; PESQ Intelligibility

1. Introduction

In recent years, synthesized speech has reached a level of quality which allows it to be integrated into many real-life applications, e.g. e-mail and SMS readers, etc. In particular, Text-to-Speech (TTS) can fruitfully be used in systems enabling interaction with an information database or a transaction server, e.g. via the telephone network.

Modern telephone networks, however, introduce a number of degradations which have to be taken into account when services are planned and developed. The type of degradation depends on the specific network under consideration. In traditional, connection-based (analogue or digital) networks, loss of loudness, frequency distortion and noise are the most significant degradations. In contrast, new types of networks (e.g. mobile or IP-based ones) introduce impairments which are perceptively different from the traditional ones. Examples are non-linear distortions from low bit-rate coding–decoding processes (codecs), overall delay due to signal processing equipment, talker echoes resulting from the delay in conjunction with acoustic or electrical reflections, or time-variant degradations when packets or frames get lost on the digital channel. A combination of all these impairments will be encountered when different networks are interconnected to form a transmission path from the service provider to the user. Thus, the whole path has to be taken into account in order to determine the overall intelligibility of the transmission network.

To quantify the intelligibility of a speech transmission chain, a number of measurement techniques have been developed during the past decades. In the subjective domain, examples are the consonant–vowel–consonant

[*] Corresponding author.

http://dx.doi.org/10.1016/j.specom.2015.04.001
0167-6393/© 2015 Elsevier B.V. All rights reserved.

(CVC) test (Steeneken, 1992a), using three-letter nonsense words in silence, and the speech reception threshold (SRT) test (Plomp and Mimpen, 1979), using short everyday sentences in noise in an adaptive procedure. Further tests are the modified rhyme test (MRT) (Fairbanks, 1958) and diagnostic rhyme test (DRT) (Voiers, 1977). The DRT and MRT tests are typical examples of so called closed response tests. In these tests subjects are offered a set of alternatives from which a selection has to be made. On the other hand, open response tests allow listeners to respond to what they think to have heard. In the objective domain, the Articulation Index (AI) (French and Steinberg, 1947), the Speech Intelligibility Index (SII) (ANSI S3.5, 1997) and the Speech Transmission Index (STI) (Steeneken and Houtgast, 1980; Steeneken, 1992b; ISO 9921, 2003; IEC 60268-16, 2003) are standardized and worldwide adopted methods for predicting the speech intelligibility for virtually any electroacoustic situation. The AI and SII (French and Steinberg, 1947; ANSI S3.5, 1997) were developed to assess speech intelligibility under conditions of additive noise and bandwidth reduction and can thus not be applied to codec and packet loss degradations as found in modern telephone networks. Basically they assess the audible parts of the speech signal and calculate a weighted average of the signal-to-noise ratios over all relevant frequency bands. From the weighted average signal-to-noise ratio the speech intelligibility can be derived. The STI method makes use of a test signal that contains spectrotemporal characteristics similar to those found in natural speech. By comparing the intensity fluctuation patterns (envelope spectra) for both the degraded output and the reference input signals, the modulation transfer function (MTF) is derived. The MTF forms the basis for quantifying how well speech information is transmitted by the transmission channel. Based on the MTF, the STI is calculated. The STI can also not be applied to codec and packet loss degradations as found in modern telephone networks, see more details in Beerends et al. (2009). Therefore, a new method for predicting the speech intelligibility in such conditions called PESQ Intelligibility (Beerends et al., 2009) based on natural speech in combination with perceptual modeling has been recently developed. An upgraded version of the PESQ Intelligibility method towards new impairments called POLQA Intelligibility, is currently being developed by Q9 of ITU-T SG12 under the work item P.OSI (series P recommendations Objective Speech Intelligibility). In fact, both PESQ Intelligibility and POLQA Intelligibility are adapted versions of particular quality prediction models, namely PESQ (Rix et al., 2002; Beerends et al., 2002; ITU, 2001) and POLQA (Beerends et al., 2013a, 2013b; ITU, 2011), for predicting the speech intelligibility.

Some work has been carried out to study the intelligibility of synthesized speech in telephony conditions and the performance of PESQ Intelligibility model for natural speech in telephony conditions. Delogu et al. (1995) evaluated the segmental intelligibility of speech transmitted over high-quality and telephone channels, using an open response test. A difference of around 10–21% was observed between the high-quality and the telephonic channel conditions for intelligibility of synthesized speech, whereas the difference for natural speech was only around 5%. Balestri et al. (1992) report on a comparative (natural vs. synthesized) speech intelligibility test for a reverse telephone directory application. Although their test was designed to be very application-specific (using representative text material, a specific prosodic structure, a real-life listener task, etc.), the results give some general indications about intelligibility differences in high-quality wideband (headphone listening) and telephonic (handset listening) conditions. All the experimental factors such as synthesized speech, bandwidth reduction, listening condition and log. PCM coding introduced perceptual degradations which also affected intelligibility. In particular, the intelligibility of natural speech was less affected by the transmission channel restriction and the handset listening (97.8–96.5%) than the synthesized speech (94.0–88.1%). The authors explain this finding by the higher cognitive demand which synthesized speech puts on the listeners.

Beerends et al. (2005) investigated to what extend PESQ (ITU-T P.862) can be used to predict the speech intelligibility with vocoders using the NATO speech intelligibility test on vocoders/noise suppressors. This database consisted of long speech files (about 3 min) containing 50 CVC words embedded in a carrier sentence. Twelve different noise conditions were used to assess the quality of 9 vocoders/noise suppressors. Bit rates of the codecs were between 1 and 5 kbit/s. The results show that PESQ provides acceptable results when used to predict the speech intelligibility. It has been also shown that some modifications of the PESQ model can increase the correlation between objective and subjective intelligibility scores from 0.86 up to 0.95. Finally, the authors have concluded that further validations are necessary in order to see if the improvements that are implemented can cope with a wide range of distortions.

Beerends et al. (2009) further exploited the idea of using a PESQ-like modeling approach in predicting subjectively obtained intelligibility scores. They focused on a large series of degradations covering band filtering, peak clipping, reverberation, noise, analog radio distortions, low bit-rate speech coding, bandwidth limitation, different types of background noise (white, babble, car), multiplicative noise and room response distortions. The results show that it is possible to develop an objective speech intelligibility measurement algorithm on the basis of PESQ despite the fact that PESQ itself shows low correlations (around 0.5) between its raw output and the subjectively obtained intelligibility scores. A simple retraining of PESQ already provides a significant improvement with a correlation of around 0.8 on untrained data. By adding advanced features the correlation between objective and subjective measurements is improved significantly and the correlation on data