Perceived Speech Quality Estimation for a Speech Streaming System via Packet Loss Network

Jin Ah Kang and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{jinari, hongkook} @gist.ac.kr

Abstract. In this paper, a speech quality assessment method is proposed to estimate perceived speech quality (PSQ) of a speech streaming system via packet loss networks. This is achieved by a simplified and low-delayed version of ITU-T Recommendation P.563. In other words, the non-intrusive assessment modules defined in ITU-T P.563 are modified to update each distortion effect using their minimum amount of speech data. After that, those effects are linearly combined once every frame using the perceptual mapping module in ITU-T P.563. The effectiveness of the proposed assessment method is then demonstrated using a speech streaming system employing the 3GPP AMR-NB speech coder. It is shown from the experiments that the proposed method gives a comparable performance on the PSQ estimation to ITU-T P.563 under packet loss conditions while significantly reducing the processing delay.

Keywords: Speech quality estimation, ITU-T Recommendation P.563, packet loss, 3GPP AMR-NB

1 Introduction

With the widespread deployment of various speech streaming services such as mobile phones and voice over IP, quality of service (QoS) has become quite critical. Accordingly, methods of speech quality assessment have been proposed to monitor the QoS of speech streaming systems [1]. Among them, ITU-T Recommendation P.563 has been popularly used to estimate perceived speech quality (PSQ) as a five-point mean opinion score (MOS) without using a reference speech signal [2]. However, ITU-T P.563 is difficult to apply to a speech streaming system that needs to monitor PSQ in real time for adaptive streaming under time-varying conditions such as packet loss rate (PLR) [3]. Thus, this paper proposes a low-delay non-intrusive perceived speech quality assessment (LD-QA) method for real-time PSQ estimation of a speech streaming system.

2 Proposed Low-Delay Perceived Speech Quality Estimation

In this section, we will briefly explain the PSQ estimation method realized in ITU-T

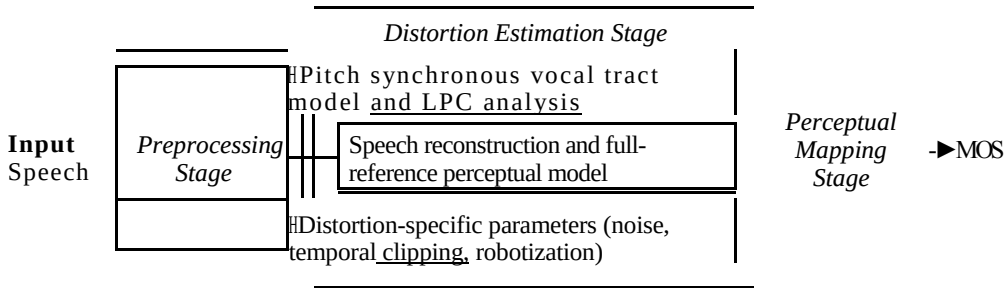

Fig. 1. Overall structure of the ITU-T P.563 model.

P.563 and then describe our proposed low-delay PSQ estimation method. Fig. 1 shows the overall structure of the ITU-T P.563 model [2], where the following three stages are processed for a given degraded input speech file due to background noise or channel distortion: preprocessing, distortion estimation, and perceptual weighting. To take into account various distortion factors during speech streaming, the model combines three processing modules in the distortion estimation stage. The first processing module models the vocal tract as a series of tubes with abnormal variations for degradation modeling and estimates the linear prediction coefficients (LPC) within a restricted range expected for a natural speech signal. The second module reconstructs a clean reference speech signal from the degraded speech signal, and then evaluates the difference between the reconstructed clean and degraded speech signals. The third module identifies and estimates specific distortions encountered in transmission channels, such as temporal clipping, robotization, and noise. On one hand, in the perceptual mapping stage, the distortion effects estimated in the distortion estimation stage are linearly combined to obtain a mean opinion score (MOS) using different weighting factors. The PSQ estimation method described so far is designed to process whole speech signals contained in an input speech file whose length should be longer than around 4 sec [2]. Therefore, this constraint causes an excessively long delay for real-time PSQ estimation.

In order to operate PSQ estimation in real time, the proposed LD-QA method modifies three processing modules in the second stage so that it can estimate the distortion effects using a minimum amount of speech data. In particular, pitch mark extraction for the vocal track analysis in the first module is modified to be processed with a speech signal of 64 ms long, and the speech reconstruction and full-reference perception model in the second module are also modified to operate once every frame. In addition, the detection process for phoneme, robotization, and temporal time clipping in the third module is modified to work using speech signals of 500 ms, 64 ms, and 1 sec long, respectively. By doing this, the perceptual quality represented as MOS is produced once every frame. Consequently, the proposed LD-QA method reduces the processing delay from around 4 sec to one frame that is typically 20 ms long.

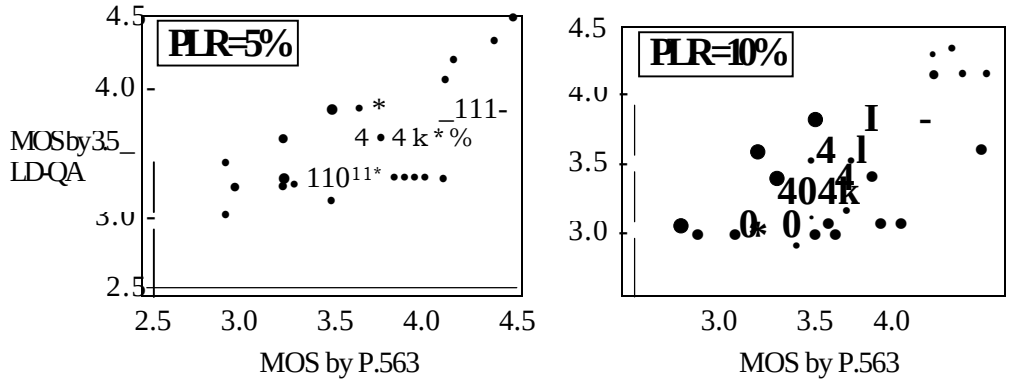

Fig. 2. Comparison of the PSQ estimated as MOS by ITU-T P.563 and LD-QA: (a) PLR = 5% and (b) PLR = 10%.

3 Performance Evaluation

In order to demonstrate the effectiveness of the LD-QA method, the estimated PSQ by the proposed method was compared with that of the ITU-T P.563 method. Toward this end, a speech streaming system was implemented using the 3GPP adaptive multi-rate-narrowband (AMR-NB) speech coder [4], where the input speech signals were sampled at 8 kHz and encoded using the AMR-NB encoder at a bit-rate of 10.2 kbit/s. In this experiment, 60 speech utterances were taken from the NTT-AT database [5]; each utterance was around 4 sec long and down-sampled from 16 to 8 kHz. For the packet loss conditions, packet loss patterns with a packet loss rate (PLR) of 5% or 10% were generated using the Gilbert-Elliot channel model defined in ITU-T Recommendation G.191 [6].

Fig. 2 shows a scattering plot of the MOS estimated by the ITU-T P.563 and LD-QA method when PLRs were 5% and 10%. Note here that the utterance-level MOS for LD-QA was obtained by averaging MOSs over all the frames in an utterance. As shown in the figure, LD-QA had correlations with the ITU-T P.563 model of 0.698 and 0.801 under 5% and 10% PLRs, respectively. Thus, we could conclude here that the performance of the LD-QA method was comparable to that of the ITU-T P.563 model, particularly at high PLR, with significantly reducing processing delay.

4 Conclusion

In this paper, we proposed a PSQ estimation method for a speech streaming system via packet loss networks. The proposed method simplified the ITU-T P.563 model so that it could operate with a much lower delay as well as provide a perceptual score once every frame. By comparing the estimated MOS of the proposed method with that of the ITU-T P.563 model, it was shown that the proposed method gave comparable performance to the ITU-T P.563 model while significantly reducing the processing delay.

Acknowledgments. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-010636).

References

1. Falk, T. H., Chan, W.-Y.: Performance study of objective speech quality measurement for modern wireless—VoIP communications. EURASIP Journal on Audio, Speech, and Music Processing, 2009, Article ID 104382, (2009) pp. 1-11.
2. Malfait, L., Bergerand, J., Kastner M.: P.563—the ITU-T standard for single-ended speech quality assessment. IEEE Transactions on Audio, Speech, and Language Processing, 14, 6, (2006) pp. 1924-1934.
3. Kang, J. A., Kim, H. K.: An adaptive packet loss recovery method based on real-time speech quality assessment and redundant speech transmission. International Journal of Innovative Computing, Information and Control, 7, 12, (2011) pp. 6773-6783.
4. 3GPP TS 26.090: Mandatory Speech Codec Speech Processing Functions; Adaptive Multi-rate (AMR) Speech Codec; Transcoding Functions, (2011).
5. NTT-AT: Multi-Lingual Speech Database for Telephonometry, (1994).
6. ITU-T Recommendation G.191: Software Tools for Speech and Audio Coding Standardization, (1996).