Speech Communication 000 (2017) 1–11

Contents lists available at ScienceDirect

Speech Communication

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

Satellite speech quality measurement model based on a combination of auditory envelope feature and link loss

Wenliang Lin*, Zhongliang Deng

Electronic Engineering Dept, Beijing University of Posts and Telecommunications, Beijing, China

ARTICLE INFO

Article history:
Received 18 February 2016
Revised 27 December 2016
Accepted 4 January 2017
Available online xxx

Keywords:
Auditory feature
Satellite mobile communication
Speech quality measurement
Signal envelope analysis

ABSTRACT

This paper focuses on Speech Quality Measurement for a satellite mobile communication system. In contrast to ground mobile communication systems, satellite speech quality measurement suffers from obvious jitter of long delays and severe satellite link losses. Auditory feature sensation and link loss measurement are most popular speech quality measurement models. However, long delay would cause the auditory feature sensation model with spreading of power spectrum, which furtherly diffuses the auditory spectrums to prevent human from hear correct responses. Nevertheless, measured error of link loss measurement would be enlarged gradually during the process of voice services. Therefore, a new speech quality measurement model based on the combination of auditory feature extraction from voice signal envelope and link loss from channel distortion is proposed. We analyze signal temporal envelope features and make power spectrum into auditory feature spectrum. The jitter of long delay and link loss are modeled as parameters to modify and compensate for the auditory feature spectrum, which also reduce the measurement distortion from transmission Doppler frequency offset to voice tone in satellite channel. A more understandable sensation estimation scores is proposed to present speech quality scores. The experimental results reveal that the new model reduces evaluation RMSE by 9.8% and is suited to the satellite mobile communications environment.

1. Introduction

Voice services are the foundation of satellite mobile communication systems. With the rapid development of communication technologies, users' requirements are improved from enabling to enjoying. A comfortable voice services would satisfy with clearness, continuity, soothing and nature. In other words, voice is not heard vaguely, interruptedly, rudely and out of tune. Measuring speech quality is an effective way to provide with judgements for the targeted voice quality (Moller and Heusdens, 2013), which guides a potential enhancement for voice services. At the case of ground mobile communication, auditory feature sensation (Kim, 2005; Hines et al., 2015; Clapham et al., 2016) and link loss measurement (Huber et al., 2014; Chappel et al., 2016; Sharma et al., 2016) are the most popular models to achieve the scores of speech quality. Auditory feature sensation is the responses to voice effected on cochlea in human's ears, which always is built as groups of filters. Link loss is more directly measurable, which is based on the equipment impairment factor method (ITU-T G.107-2015). It takes all qualities loss from different processes of whole voice services into consideration, such as voice generation, voice transition and voice hearing. However, the obvious jitter of long delays and severe satellite link losses would degrade the performance of speech quality measurement. Fig. 1 tries to demonstrate those performance degradations under satellite links. For auditory feature sensation model, long delay would cause the spreading of power spectrum, which furtherly diffuses the auditory spectrums to prevent human from hearing correct responses. Nevertheless, focused on link loss measurement, we figure out that measured error would be enlarged gradually during the process of voice services, which is called error avalanche. Once measured error increases, the total measured error would be much bigger than the former contributed.

Global efforts are made to achieve more correct scores for speech quality measurements. The ANIQUE (Auditory Model for Single-Ended Speech Quality Estimation) model (Kim, 2005) was proposed to extract voice auditory features from temporal envelope of voice signals. It was constructed by a mathematical parameter of voice ANR (Articulation to Non-Articulation Ratio), which was the percentage of effective power taken in a voice signal. However, power and frequency are not the whole effects on auditory features, tone also plays important role of speech quality. ANIQUE would miss some key features to reproduce the experi-

* Corresponding author.
E-mail address: Charterlin@163.com (W. Lin).

http://dx.doi.org/10.1016/j.specom.2017.01.001
0167-6393/

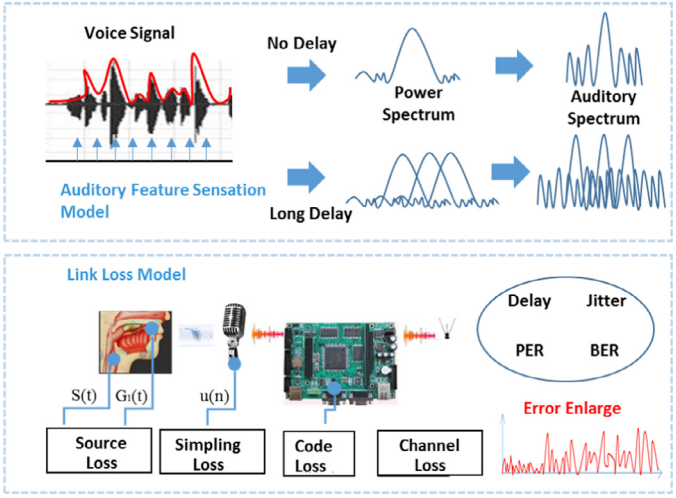

Fig. 1. The poor measured performance under satellite links.

ence of good voice, resulting in measurement distortion. A speech quality measurement scheme based on the GM (Gaussian Mixture) (Huber et al., 2014) was attempted to measure the loss between received signal and original signal. Combined with link parametric, it provided with PER (Packet Error Rate) to predict spectrum degradation. Ignored of auditory features, GM model could not reach a human understandable level to guide exact quality scores. Since, environment also made influence on voice capture and hearing (Li et al., 2006), the CSCA (Combined Computational Auditory Scene Analysis) model (Dubey, 2015) provided with hybrid ambient noise to modify auditory distortion to compensate for measurement. Nevertheless, lack of self-adaptive sensation, CSCA model cannot suffer from serious jitter of long delays so estimation performance degraded. The link loss model combined with signaling transition situation (Chen et al., 2011) tried to detect signaling attenuation to get further link loss parameters, but rude compensation failed to reach smooth auditory experience. Uncertainty analysis (Deng et al., 2015; Paglierani and Petri, 2009) was an interesting approach to limiting measurement error range of speech quality, but the measurement model was still restricted by upper limitation of auditory feature sensation model.

Above all, auditory feature sensation model can achieve nature feature of voice, which present real feeling affected on human sensation. But spreading of power spectrum by jitter of long delays impedes more correct speech quality measurement (Falk and Chan, 2008). What's more, macro scores always can't be divided into some detailed parameters with different scales. We have no direct improvement suggestions acquired from auditory feature quality measurement. The link loss model is supposed to test some parametric from communication channels. It seems more scientific and reasonable. After all, voice still has to be changed into human's affection on hearing sensation to optimize estimations. Coming very naturally, combined auditory feature sensation model and link loss model may contribute an ideal scheme to balance voice nature feature presentation and link distortion to reduce measurement performance degradation.

Therefore, a new speech quality measurement model is proposed based on the combination of auditory feature extraction from voice signal envelope and link loss from channel distortion. We fully use nature features on human cochlea to describe the comfort from voice, while the equipment impairment measurements are also considered. The voice quality in satellite link is also concerned. Long delay and high attenuation make great influences on voice and voice quality measurement. They come from different processes. We analyze signal temporal envelope features and make power spectrum into auditory feature spectrum. The jitter of long delay and link loss are modeled as parameters to modify and compensate for the auditory feature spectrum. The jitter of long delay and link loss are modeled as parameters to modify and compensate for the auditory feature spectrum, which also reduce the measurement distortion from transmission Doppler frequency offset to voice tone in satellite channel. A more understandable sensation estimation scores is proposed to present speech quality scores. At last, new model successfully achieve more precise results and closer to the voice nature.

2. New model combined envelope feature and link loss

The satellite speech quality measurement model based on the combination of envelope feature and link loss comprises three main parts: the voice temporal envelope features extraction and delay pre-process, the equivalent auditory spectrum with link loss compensation and the sensation estimation scores. The framework of the new model is presented in Fig. 2.

As Fig. 2 shown, first, the chosen reference voice signal $X(t)$ is transmitted by the targeted satellite mobile communication system. The received and decoded signal is the degraded signal $Y(t)$. Then, the total level normal factor is calculated from the mean power of $X(t)$ and $Y(t)$, then limited by the base SNR (Signal to Noise Ratio) parameter from the E-model R_0 (ITU-T G.107-2015) to normalize the levels of both signals. Based on the SPL (Sound Pressure Level), both signals travel through block filter with linear response with the features of the satellite terminal speaker characteristic and the equipment loss parameter I_{e-eff} (ITU-T G.107-2015). Second, both filtered signals are extracted from the voice temporal envelope to obtain the VAD (Voice Activity Detection). The delay loss parameter is set as the reference point, and the delay between two voice sub-frames is aligned to achieve precise time alignment. Instead of the Bark spectrum (Moller and Heusdens, 2013), the TMTF (Temporal modulation transfer function) is designed to transform the frequency spectrum to the temporal envelope feature spectrum. Third, the interference density and masking value are calculated by the optimized loudness spectrum generated with the control factor. Finally, we obtain the aggressive speech quality sensation scores.

2.1. Voice temporal envelope feature extraction and delay pre-processing

The chosen reference voice signal $X(t)$ and degraded signal $Y(t)$ must operate the voice temporal envelope feature extraction and delay pre-processing. The main procedure comprises three steps: signal level normalization, linear response block filtering with the satellite terminal speaker characteristic and equipment loss, and temporal envelope feature extraction and smooth time alignment.

2.1.1. Voice signal level normalization

For more precise level normalization, the base SNR parameter from the E-model R_0 (ITU-T G.107-2015) is used:

$$R_0 = 15 - 1.5(SLR - N_0) \tag{1}$$

Then, the total normalized factor A from signal voltage level is calculated by dividing the 80 dB satellite speaker SPL by the mean power of $X(t)$ and $Y(t)$, as well as R_0:

$$A = \frac{80\text{dB } SPL}{\frac{1}{2}\sqrt{10\lg(\int_{-\infty}^{+\infty} X^2(t)dt + 10\lg\left(\int_{-\infty}^{+\infty} Y^2(t)dt\right) - N_0}} \tag{2}$$

Taking A as the reference level, the power of both signals are adjusted to the same normalization level to provide a relative power range. SPL is the local pressure deviation from the ambient

Please cite this article as: W. Lin, Z. Deng, Satellite speech quality measurement model based on a combination of auditory envelope feature and link loss, Speech Communication (2017), http://dx.doi.org/10.1016/j.specom.2017.01.001