PERCEPTUAL MODEL FOR NON-INTRUSIVE SPEECH QUALITY ASSESSMENT

Doh-Suk Kim, Ahmed Tarraf

Lucent Technologies
67 Whippany Road, Whippany, NJ 07981, USA
Email: dsk@lucent.com, tarraf@lucent.com

ABSTRACT

In order to estimate the quality of degraded speech processed by communication networks, conventional objective speech quality assessment methods require source speech signal, which has been applied to the networks, as well as the processed speech. This paper presents a new paradigm in objective speech quality assessment. In contrast to previous objective models, the proposed Auditory Non-Intrusive QUality Estimation (ANIQUE) model estimates the quality of speech without using the source speech information at all. ANIQUE is a perceptual model which simulates the functional role of human auditory system at both peripheral level and central auditory level. The performance of ANIQUE is demonstrated using 19 different subjective MOS databases.

1. INTRODUCTION

Modern speech communication networks are becoming more and more complex. In addition to existing traditional Public Switched Telephone Network (PSTN), various types of new networks such as wireless and Voice over Internet Protocol (VoIP) are being used in daily life. As new technologies are emerged and converged to existing telephone network infrastructure, we are facing many factors degrading perceived quality of speech, and the measurement of speech quality becomes very important in the sense that it's not only a starting point to improve quality of service (QoS) but also a maintenance tool for quality satisfaction.

The most reliable way to measure the quality of speech is to perform very well-controlled subjective speech quality assessment tests. In these tests a large number of subjects listen to hundreds of short speech utterances processed by the system under test, and rate its performance as e.g., a five-point scale. The average rating is commonly referred to as Mean Opinion Score (MOS) [1]. Obviously, these tests are expensive both in time and cost, and difficult to reproduce. Thus, it is desirable to have an objective method which can reflect subjective ratings on speech signal in reliable manner, and many objective measures were proposed and are being used in real applications [2, 3, 4, 5]. These conventional models are basically intrusive methods, which rely on a distance metric typically intended to model the functional role of human auditory perception. Thus in addition to the processed speech signal, we need to have the source speech information which was used as an input to the system under test. This is illustrated in Fig. 1 (a).

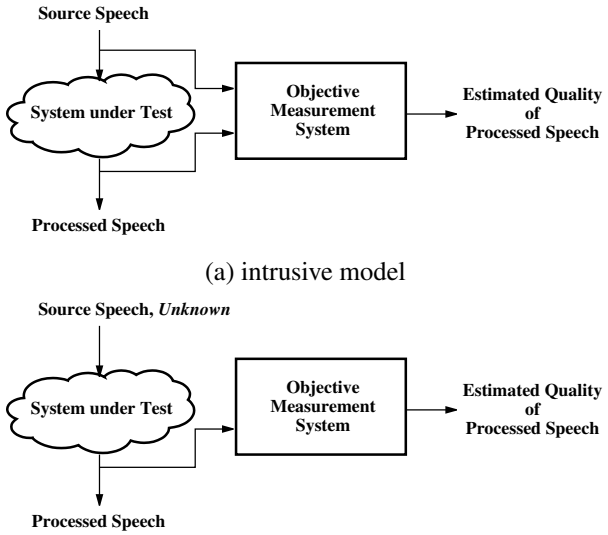

Fig. 1. Intrusive and non-intrusive speech quality estimation models.

However in real subjective MOS tests, speech signal is not presented to listeners in pair-wise comparison together with its source speech. And human listeners do not create a hypothesized source speech from processed speech, unlike the common postulate that conventional intrusive models take. Moreover, there are many scenarios in the application of objective measures where source speech information is not available. A non-intrusive method is a challenging paradigm in objective speech quality assessment in the sense that it requires only processed speech signal without having source speech information at all, as shown in Fig. 1 (b) where the source speech is unknown. In this paper a perceptual model, Auditory Non-Intrusive QUality Estimation

(ANIQUE), is proposed for objective non-intrusive speech quality assessment. The proposed model is based on the modeling of functional roles of human auditory system both at peripheral level and at central level to estimate the quality of speech in non-intrusive manner.

2. ANIQUE MODEL

Fig. 2 shows the block diagram of the proposed ANIQUE model, and the brief description of each block is provided in the following.

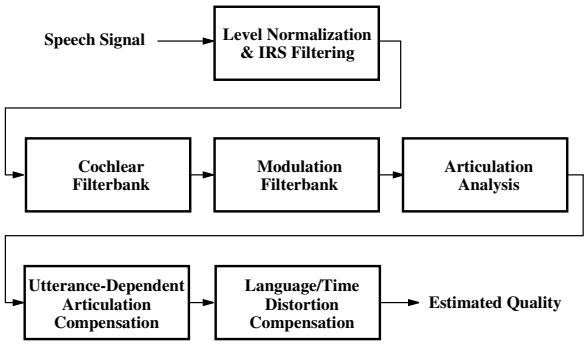

Fig. 2. Block diagram of ANIQUE model.

2.1. Level Normalization and IRS Filtering

The level of speech signal is first normalized to -26 dBov using P.56 speech voltmeter [6]. Then, Intermediate Reference System (IRS) receive filter is applied to reflect the characteristics of handsets used in subjective listening tests [7].

2.2. Cochlear Filterbank

Simulating the first stage of human auditory system, the normalized and IRS-filtered speech signal, $s(n)$, is filtered by a bank of critical-band filters, $h_k(n)$, $k = 1, 2, \ldots, N_{cb}$, where $h_k(n)$ is the impulse response of the k-th critical-band filter and N_{cb} denotes the number of filter channels. The critical band signal at the k-th channel is represented as

$$s_k(n) = s(n) * h_k(n). \tag{1}$$

The characteristic frequency of the filters in cochlear filterbank ranges from 125 Hz to 3500 Hz, and the bandwidth of each cochlear filter is characterized by equivalent rectangular bandwidth (ERB) [8].

For each critical band, the temporal envelope is obtained as

$$\gamma_k(n) = \sqrt{s_k^2(n) + \hat{s}_k^2(n)} \tag{2}$$

and the instantaneous phase is represented as

$$\phi_k(n) = \arctan \frac{\hat{s}_k(n)}{s_k(n)} \tag{3}$$

where $\hat{s}_k(n)$ is the Hilbert transform of $s_k(n)$. Now we can express $s_k(n)$ in terms of its temporal envelope and carrier as

$$s_k(n) = \gamma_k(n) \cos \phi_k(n). \tag{4}$$

Fig. 3 illustrates how temporal envelope is represented in speech signal. The top panel (a) shows an example of a female speech segment /*a*m/ passed through a critical-band filter centered at 1050 Hz. The bottom plot (b) depicts the temporal envelope of (a). The temporal envelope shows modulation components caused by glottal excitation at 184 Hz (pitch) and the movement of human articulatory system at 2 $\sim$ 30 Hz.

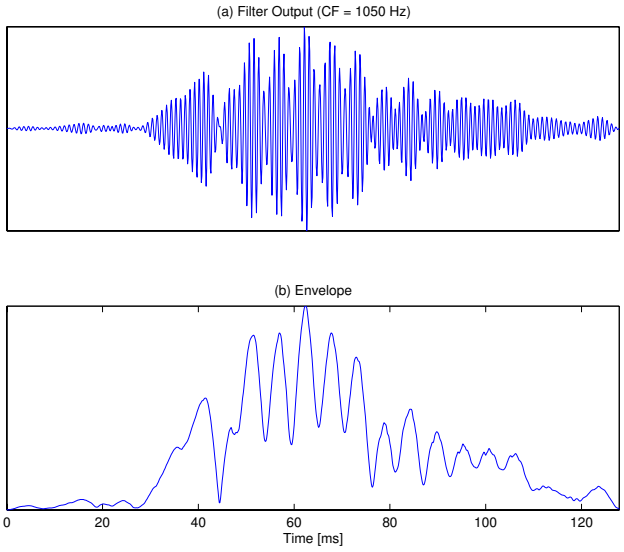

Fig. 3. Example of envelope signal and its modulation spectrum: (a) 128 ms-long output of a critical-band filter centered at 1050 Hz; (b) the temporal envelope of (a).

The decomposition of speech signal into its temporal envelope and carrier provides useful insights in speech perception, because temporal envelope is known to be relevant to many perceptual attributes of speech, such as intelligibility and quality. In terms of speech quality, it's worth noting de Boer's 'phase rule', which states that the quality of a sound is unchanged if the phases of the components are shifted by a constant amount and/or amount that are linearly dependent on the frequency of the components [9]. Considering the phase changes obeying this rule give no change in the temporal envelope of the signal, the change of sound quality is directly related to the change of envelope.

2.3. Modulation Filterbank and Articulation Analysis

The mechanism of human sensitivity to the temporal envelope of stimuli is an interesting topic in psychophysics, and Dau et. al. proposed an auditory model in which a bank of modulation detectors is employed to describe the modulation detection and modulation masking data obtained in psychophysical experiments in 1997 [10, 11]. Neurophysiological studies support this idea and showed the higher auditory pathway is organized as a hierarchical filterbank [12].

In the proposed ANIQUE model, the higher level of auditory pathway is modeled by the modulation filterbank and following articulation analysis. For each critical band, the frame signal of temporal envelope is obtained by multiplying $\gamma_k(n)$ to 256 ms Hamming window, which is shifted by 64 ms every frame. Fourier transform is then performed on the frame envelope, resulting in "modulation spectrum":

$$\Gamma_k(m, f) = \mathcal{F}\left\{\gamma_k(m; n)\right\}, \tag{5}$$

where $\gamma_k(m; n)$ is the m-th frame signal of $\gamma_k(n)$ and f represents modulation frequency.

The modulation spectrum is grouped into M bands by a modulation filterbank $\{W(i, f)|i = 1, 2, \ldots, M\}$, resulting in modulation band power:

$$\Psi_k(m, i) = \sum_f \Gamma_k^2(m, f) W(i, f)^2. \tag{6}$$

The modulation filterbank is implemented and applied to every cochlear channel in modulation frequency domain, where the quality factor of each filter is set 2.

In ANIQUE model, it is hypothesized that human listeners determine the quality of speech by making use of internal knowledge in modulation spectral domain - the higher level of auditory pathway segregates signal components produced by human speech production system from the others such as coding distortions and noise. For the mathematical formulation, articulation-to-nonarticulation ratio (ANR) at the k-th cochlear channel is defined as

$$\Lambda_k(m) \quad = \quad \frac{\Psi_{k,A}(m)}{\Psi_{k,N}(m)}. \tag{7}$$

Here, the numerator is the average articulation power taken from the first four modulation band powers, i.e., $i = 1, \ldots, 4$ in Eq. (6), to cover the frequency range of human articulation motor system (2 – 30 Hz). The denominator is the average nonarticulation power to reflect the effect of distortions which cannot be generated by human articulation systems. The frequency range to calculate average nonarticulation power is set different for different critical bands, based on Ghitza's investigation on the upper cutoff frequency of the critical-band envelope detectors [13]. In his psychophysical experiments, it was shown that minimum bandwidth of the envelope information for a given auditory channel is roughly the half of critical bandwidth in order to preserve speech quality, which implies the modulation frequency components of temporal envelope is relevant to the perception of speech quality only up to the half of critical bandwidth.

2.4. Frequency and Time Aggregation

The ANRs for all cochlear channels are accumulated to yield the frame quality as

$$\nu_s(m) = \left[\sum_{k=1}^{N_{cb}} \Lambda_k^p(m)\right]^q, \tag{8}$$

where $\Lambda_k(m)$ is the ANR of the k-th cochlear channel at the m-th frame, and p and q are empirically determined values.

Based on the value of the dc-component of modulation power spectrum, i.e., $\Gamma_k^2(m, 0)$, each frame is categorized into three events: loud event (LE), faint event (FE), and inaudible event (IE).

Then the overall speech quality is obtained as

$$\begin{aligned} Q_s &= \alpha Q_{s,LE} + (1-\alpha) Q_{s,FE} \\ &= \alpha \left[\sum_{m \in LE} \nu_s^3(m)\right]^{\frac{1}{3}} + (1-\alpha) \left[\sum_{m \in FE} \nu_s^3(m)\right]^{\frac{1}{3}} \end{aligned} \tag{9}$$

where α is a weighting factor for loud event.

2.5. Utterance-Dependent Articulation Compensation

Depending on phonetic contents, speaking styles, and individual speaker differences, different utterances with same subjective quality can have different ANR distributions, and it is necessary to compensate this effect. Considering Modulated Noise Reference Units (MNRU) [14] with several different SNRs are used as anchor points in subjective tests for introducing controlled degradation to speech signals, the MNRU with very low SNR is produced from $s(n)$ as

$$q(n) = s(n)[1 + 10^{-SNR/20} d(n)], \tag{10}$$

where $d(n)$ is random noise and SNR is the ratio of speech power to modulated noise power in dB.

MNRU signal $q(n)$ is then processed same way as $s(n)$, and Q_q, the quality of the signal $q(n)$, is obtained similar to Eq. (9). Q_q is regarded as the minimum unit quality for the speech signal $s(n)$, and the compensated speech quality is obtained as

$$\tilde{Q}[s(n)] = Q_s / Q_q. \tag{11}$$

2.6. Language / Time Distortion Compensation

In typical subjective listening tests, native listeners are recruited as subjects, i.e., the language in speech material presented to listeners is the mother tongue of listeners. Since some samples of time-related network distortions (such as time-clipping) of speech can make damages to its language contents, their impacts on speech quality can be more significant to native listeners than to non-native listeners. In objective method perspective, it is necessary to employ a mechanism which can reflect the language effect to the quality estimation. In ANIQUE model, time-related distortions are detected by using time-derivatives of envelope, and corresponding quality is adjusted to reflect their impacts on the quality of speech rated by native listeners.

3. EXPERIMENTAL RESULTS

An experimental evaluation was performed using 19 subjective MOS databases, each of which consists of hundreds of speech samples and their associated MOS values. These databases cover wide range of telecommunication applications – standard and nonstandard speech codecs, transcoding, channel errors, packet loss and its concealment, environmental noise at sending side, time-varying delay, VoIP, and so forth.

The most common metric for evaluating the performance of objective speech quality estimation methods is the correlation coefficient between subjective and objective values. The proposed ANIQUE model shows average correlation of 0.8546 over all 19 databases (per-condition correlation after 3rd order monotonic polynomial regression). For the same task, ITU-T recommendation P.862 (PESQ) shows the average correlation of 0.932. Although the performance of ANIQUE is lower than that of PESQ, the result ANIQUE demonstrated is quite promising considering the fact that PESQ uses the information on source speech as well as processed speech.

4. CONCLUSIONS

This paper presents a new paradigm in objective speech quality assessment. The proposed ANIQUE model is based on the functional role of human auditory system in judging the quality of speech, and consists of critical-band filters, modulation filterbank, articulation analysis, and compensation stages. In contrast to conventional intrusive objective speech quality methods, ANIQUE model estimates the quality of speech without using the source speech information, and this methodology is more analogous to real subjective MOS tests.

5. REFERENCES

[1] ITU-T Recommendation P.800, *Methods for objective and subjective assessment of quality*, 1996.

[2] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, March 1994.

[3] ITU-T Recommendation P.861, *Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Codecs*, Geneva, 1996.

[4] S. Voran, "Objective estimation of perceived speech quality, Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 4, 1999.

[5] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, Geneva, 2001.

[6] ITU-T Recommendation P.56, *Objective measurement of active speech level*, 1993.

[7] ITU-T Recommendation P.48, *Specification for an intermediate reference system*, 1988.

[8] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–108, 1990.

[9] E. de Boer, "A note on phase distortion and hearing," *Acoustica*, vol. 11, pp. 182–184, 1961.

[10] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. America*, vol. 102, pp. 2892–2905, 1997.

[11] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. America*, vol. 102, pp. 2906–2919, 1997.

[12] A. Giraud, C. Lorenzi, J. Ashburner, J. Wable, I. Johnstude, R. Frackowiak, and A. Kleinschmidt, "Representation of the temporal envelope of sounds in the human brain," *J. Physiology*, pp. 1588–1598, 2000.

[13] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. America*, vol. 110, no. 3, pp. 1628–1640, Sep 2001.

[14] ITU-T Recommendation P.810, *Modulated Noise Reference Unit (MNRU)*, Feb. 1996.