ENHANCED 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 this paper, we propose a novel model for estimating the quality of speech without the reference speech information. The proposed auditory non-intrusive quality estimation plus (ANIQUE+) model is a perceptual model simulating the functional role of human auditory system, and employs improved modeling of quality estimation by statistical learning methods. Experimental evaluation demonstrated that the performance of the ANIQUE+ model is significantly superior to that of the current ITU-T standard recommendation P.563 on 34 different subjective mean opinion score (MOS) databases – the averaged correlation between subjective and objective quality scores is about 0.97 for ANIQUE+, whereas P.563 shows 0.87 averaged correlation.

1. INTRODUCTION

Non-intrusive estimation of speech quality is a challenging problem in that it estimates the quality of speech transmitted over telecommunication networks without using the reference speech information, in contrast to intrusive models such as ITU-T P.862 PESQ [1] which compare reference and degraded speech signals in estimating the quality of degraded speech. Recently in the ITU-T, the standard recommendation P.563 has been adopted for non-intrusive estimation of speech quality as the need for monitoring the speech quality of in-service networks is rapidly growing where no reference speech signal is available [2].

However, ITU-T P.563 model demonstrates very limited performance. The averaged correlation between subjective mean opinion score (MOS) and objective quality estimated by P.563 is about 0.88 even for the 24 known MOS test databases used in the development of the model, whereas ITU-T P.862 shows 0.93 correlation for the same task [3]. An experiment has also been reported that the performance of P.563 is quite unsatisfactory for some unknown MOS tests (MOS test data not included in the model development) containing selectable mode vocoder (SMV) conditions (the correlation is as low as 0.7) [4]. In order to use a non-intrusive quality estimation model in real applications, higher performance closer to the performance level of P.862 is required.

The authors proposed auditory non-intrusive quality estimation (ANIQUE) model previously, which is based on the temporal envelope representation of speech motivated by the functional roles of human auditory system [5, 6]. In this paper, we presents an enhanced version of ANIQUE model, ANIQUE+, in which improved modeling of quality estimation is employed based on statistical learning method. Experimental evaluation demonstrates the performance of ANIQUE+ model is significantly higher than the ITU-T P.563.

2. ANIQUE+ MODEL

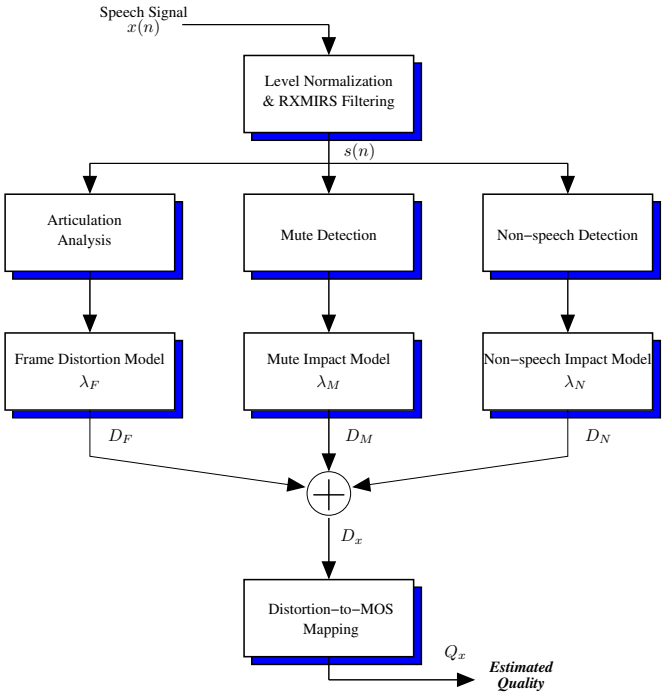

Fig. 1. Block diagram of the ANIQUE+ model.

Fig. 1 shows the overall block diagram of the proposed ANIQUE+ model. The overall objective distortion D_x, tak-

ing the value between 0 and 1, of the speech signal $x(n)$ is estimated by the sum of the overall frame distortion D_F, the mute distortion D_M and the non-speech distortion D_N:

$$D_x = D_F + D_M + D_N. \quad (1)$$

The distortion D_x is then mapped onto subjective MOS scale to yield objective speech quality Q_x:

$$Q_x = -3.5 \min(D_x, 1) + 4.5 \quad (2)$$

assuming the maximum and minimum value of quality are 4.5 and 1.0, respectively.

2.1. Level Normalization & Receive-Side Modified IRS Filtering

The level of speech signal is first normalized to -26 dBov using the P.56 speech voltmeter [7]. Then the modified intermediate reference system (IRS) receive filter is applied to reflect the characteristics of the handset used in listening tests [8].

2.2. Cochlear Filterbank and Temporal Envelope

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 critical bands. The critical band signal at the k-th channel is represented as

$$s_k(n) = s(n) * h_k(n). \quad (3)$$

The characteristic frequency of the filters in cochlear filterbank ranges from 125 Hz to 3500 Hz, and the bandwidth of each critical-band filter is characterized by equivalent rectangular bandwidth (ERB) [9]:

$$ERB_k = F_k/Q_{ear} + B_{min} \quad (4)$$

where F_k is the characteristic frequency of the k-th critical-band filter in Hertz, and Q_{ear} and B_{min} are set to 9.26449 and 24.7, respectively.

2.3. Modulation Filterbank and Analysis

In addition to the existence of frequency selectivity at peripheral level, it is believed that there is a set of modulation detectors, each of which is tuned to a specific modulation frequency, at the central level of the auditory system [10]. This idea was adopted in the ANIQUE model [5, 6], and then in the proposed ANIQUE+ model. For each critical band, the temporal envelope of $s_k(n)$ is obtained by

$$\gamma_k(n) = \sqrt{s_k^2(n) + \hat{s}_k^2(n)} \quad (5)$$

where $\hat{s}_k(n)$ is the Hilbert transform of $s_k(n)$. The temporal envelope is then multiplied by the 256 ms Hamming window, which is shifted by 64 ms every frame, in order to obtain $\gamma_k(m; n)$, which is the temporal envelope for the k-th critical band at the m-th frame. The modulation spectrum for each critical band is then estimated by Fourier transform as

$$\Gamma_k(m, f) = |\mathcal{F}\{\gamma_k(m; n)\}| \quad (6)$$

where f represents modulation frequency.

The modulation spectrum is grouped into M bands by a modulation filterbank $\{W(i, f)|i = 1, 2, \ldots, M\}$, and one can obtain modulation band power as

$$\Psi_k(m, i) = \int \Gamma_k^2(m, f) W^2(i, f) df, \quad (7)$$

where the modulation filterbank is a set of M equal-Q filters (Q = 2) implemented in modulation frequency domain, and its frequency response can be found in [6].

In [6], the articulation-to-nonarticulation ratio (ANR) for the k-th critical band is defined as

$$\Lambda_k(m) = \frac{\Psi_{k,A}(m)}{\Psi_{k,N}(m)} \quad (8)$$

where the numerator is the average articulation power reflecting signal components relevant to natural human speech, and the denominator is the average nonarticulation power representing perceptually annoying distortions produced at the rates beyond the speed of human articulation systems [6]. The ANR is aggregated across all the critical bands to compute the frame quality.

The introduction of ANR provides simple yet effective method for frame quality estimation, but it simplifies the human quality perception and ignores the interaction across critical bands, which is believed to exist especially at the higher level of auditory pathway. As the detailed mechanism of quality perception by human listeners is not known yet, data-driven approach is employed in the proposed ANIQUE+ in which the objective model learns the relationship between speech signals and their associated quality ratings. Instead of using ANR, the feature vector for frame distortion model in the ANIQUE+ consists of the articulation power, nonarticulation power and critical band power, resulting in a $(3N_{cb})$-dimensional vector at the m-th time frame:

$$\mathbf{\Xi}(m) = [\mathbf{\Psi}_A(m); \mathbf{\Psi}_N(m); \log \mathbf{\Gamma}(m, 0)] \quad (9)$$

where $\mathbf{\Psi}_A(m)$, $\mathbf{\Psi}_N(m)$, and $\mathbf{\Gamma}(m, 0)$ are vector representations of $\Psi_{k,A}(m)$, $\Psi_{k,N}(m)$, $\Gamma_k(m, 0)$, respectively.

2.4. Frame Distortion Model

In the frame distortion model, the overall frame distortion D_F is modeled as

$$D_F = D_S + D_B. \quad (10)$$

Here, D_S is the distortion in speech obtained by accumulating frame distortions for active speech over time and then normalizing by the total number of active speech frames T_S as

$$D_S = \frac{1}{T_S} \sum_{m \in \mathcal{S}} \chi(m) \tag{11}$$

where $\chi(m)$ is the output of frame distortion model ranging from 0 to 1 at the m-th frame. D_B is the audible distortion in background noise and is estimated as

$$D_B = \frac{1}{T_B} \sum_{m \in \mathcal{B}} \{\alpha_F (P_{env}(m) - P_{th}) + \beta_F\} \chi(m) \tag{12}$$

where $P_{env}(m)$ is the DC-value of modulation power spectrum at the m-th frame, P_{th} is the threshold for audible background noise, T_B is the number of frames for background noise, and α_F and β_F are weighting factors.

The frame distortion model used in the ANIQUE+, λ_F, is the multi-layer perceptron with one hidden layer, and its output is expressed as

$$\chi(m) = g(\sum_j W_j g(\sum_k w_{jk} \xi_k(m))) \tag{13}$$

where $\xi_k(m)$ is the k-th element of input feature vector $\boldsymbol{\Xi}(m)$, w_{jk} and W_j are synaptic weights for the input and hidden layer, respectively, and $g(x)$ is the nonlinear sigmoid function. Synaptic weights are obtained by error back-propagation learning [11]. Fig. 2 illustrates how the speech distortion D_S is estimated.

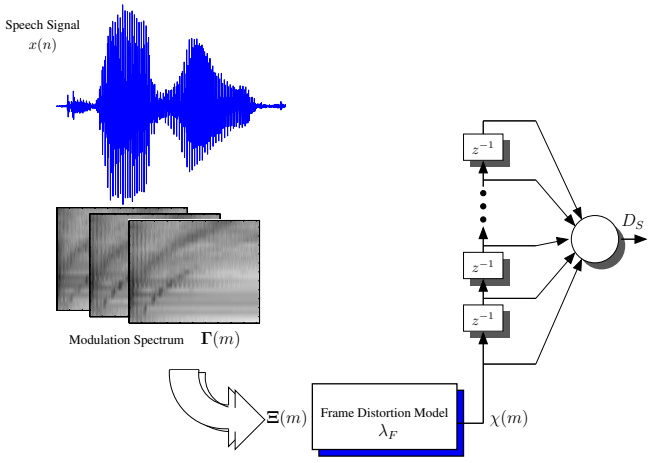

Fig. 2. Frame distortion estimation model.

2.5. Mute Detection and Impact Model

Interruption such as mutes is the one of most common distortions observed as packet loss or frame erasure in modern wireless and VoIP networks. The purpose of the model in this section is to detect unnatural mutes in speech signals and estimate their impact on perceived quality.

2.5.1. Detection of Unnatural Stops

At every possible candidate time frame for the beginning of mute, l_M, where the frame energy drops abruptly, a feature vector is extracted for two time instances, l_M and 15 ms prior to l_M, with the analysis length of 30 msec. The feature vector includes the 12-th order Mel-Frequency Cepstral Coefficients (MFCC), and voicing factor, which indicates how much periodic components a segment of speech contains. A neural network detector (multi-layer perceptron) with one output neuron was trained to detect unnatural abrupt stops on training database.

2.5.2. Detection of Unnatural Starts

When frames start to be erased during silence even before a speech activity starts, the beginning of mute cannot be detected. In this case, only the end of mute exists and is termed 'unnatural abrupt start'. Similar to abrupt stop detection, a feature vector is extracted for two time instances, l_M and 15 ms after l_M, where l_M is the candidate time frame of the beginning of unnatural abrupt start. The feature vector includes spectral centroid defined as

$$s = \frac{\sum_k k|X(k)|}{\sum_k |X(k)|} \tag{14}$$

where $|X(k)|$ is the FFT magnitude of a speech frame, as well as MFCCs and voicing factors. A neural network detector was trained to detect unnatural abrupt starts on training database.

2.5.3. Impact of Mutes

Recent experiments revealed that humans can assess the quality of speech continuously over time and there's some recency effects in perceived overall quality [12]. This is related to biological short-term memory and means that recent events can play more role than past ones. In the proposed ANIQUE+, the impact of mutes is modeled as the combination of abrupt instantaneous distortion followed by decays simulating short-term memory effects.

2.6. Non-Speech Detection and Impact Model

This module detects very annoying non-speech activities that may occur when bit information within a packet or frame is distorted during transmission but not detected at the speech decoder side, for example. In the ANIQUE+ model, the time-derivative of frame power is used to detect non-speech activities and its contribution to objective distortion is estimated proportional to the energy of non-speech activities.

3. EXPERIMENTAL RESULTS

The proposed ANIQUE+ model was developed using the 24 MOS test databases (speech files and their associated MOS

Table 1. Performance of the ANIQUE+ model on two different data sets. For the performance metric, per-condition correlation coefficient (ρ) and root mean squared error (RMSE) of MOS after 3rd-order monotonic polynomial regression are used.

(a) 24 known MOS tests

	ANIQUE+	P.563	P.862
ρ	0.9762	0.8787	0.9300
RMSE	0.1514	0.3714	0.2707

(b) 10 unknown MOS tests

	ANIQUE+	P.563	P.862
ρ	0.9388	0.8480	0.9528
RMSE	0.2258	0.3045	0.1795

values), which are the same data used in the ITU-T P.563 selection phase and consist of about 16 hours of speech covering wide range of telecommunication applications. To demonstrate the validity of the model further, 10 unknown MOS test databases of 34 hours os speech were used which have never been utilized in training of the model. They include SMV characterization databases, low bit rate CELP codec tests, G.728 characterization tests, and VoIP conditions.

Table 1 shows the performance of the ANIQUE+ model in comparison with ITU-T standard recommendations, P.563 and P.862. On the 24 known MOS tests, which were used in the training of both ANIQUE+ and P.563, ANIQUE+ shows significantly higher performance than P.563 and even higher than P.862, which is an intrusive model utilizing reference source speech as well as degraded speech signals. On the 10 unknown MOS test, the performance of ANIQUE+ is much superior to that of P.563, and the performance gap between intrusive and non-intrusive models has been extensively reduced. The averaged correlation across all 34 databases is 0.97 for ANIQUE+ and 0.87 for P.563.

4. CONCLUSIONS

This paper presents an enhanced method in objective speech quality assessment. The proposed ANIQUE+ model is based on the functional role of human auditory system in rating the quality of speech, and consists of critical-band filters, modulation filterbank, articulation analysis, and three models for estimating perceptual distortion in speech signals. In contrast to previous ANIQUE model, statistical learning based on training data is employed instead of establishing models based on insufficient knowledge and assumptions, resulting in significantly better performance than the current ITU-T standard recommendation P.563 not only for the data used in training but also for unknown data sets.

5. REFERENCES

[1] 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.

[2] ITU-T Recommendation P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications*, Geneva, 2004.

[3] ITU-T COM 12-D208E, *Performance of the P.SEAM reference model*, Geneva, March 2004.

[4] D. -S. Kim and A. Tarraf, *Single-Ended Model for Objective Speech Quality Estimation - Its Performance on SMV Characterization Database*, 3GPP2 TSG-C WG1 SWG 1 Meeting, Dec. 2004.

[5] D.-S. Kim and A. Tarraf, "Perceptual model for non-intrusive speech quality assessment," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, Montreal, Canada, May 2004, pp. 1060–1063.

[6] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

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

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

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

[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] D. Rumelhart, G. Hinton, and R. Williams, *Learning Internal Representation by Error Propagation*. MIT Press, 1986.

[12] M. Hansen and B. Kollmeier, "Continuous assessment of time-varying speech quality," *J. Acoust. Soc. America*, vol. 106, no. 5, pp. 2888–2899, 1999.