A HYBRID SIGNAL-AND-LINK-PARAMETRIC APPROACH TO SINGLE-ENDED QUALITY MEASUREMENT OF PACKETIZED SPEECH

Tiago H. Falk, Hua Yuan, and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
E-mail: {falkt, yuanh, chan}@ee.queensu.ca

ABSTRACT

A hybrid signal-and-link-parametric approach to single-ended quality measurement of packetized speech is proposed. Transmission link parameters are used to determine a base quality for the test signal. The base quality is adjusted by degradation factors calculated from perceptual features extracted from the test signal. The degradation factors are based on Kullback-Leibler distances between a parametric model trained online for the extracted features and reference models of normative speech behavior. The proposed method overcomes the limitations of pure link parametric and pure signal-based methods.

Index Terms— Quality measurement, VoIP, packet loss concealment, Kullback-Leibler distance.

1. INTRODUCTION

Voice over internet protocol (VoIP) uses packet transmission of speech over the Internet. Current IP networks are optimized for data communications where variable losses and delay are not critical since retransmission can be performed. With voice communications, however, retransmission of missing packets is not a viable option. Packet losses can occur due to network delay, network congestion or network errors. Lost packets can degrade the quality of the transmitted speech signal considerably. Packet loss concealment (PLC) algorithms are used to replace lost packets and to improve speech quality.

The performance of different PLC algorithms can be assessed via subjective listening tests, such as the mean opinion score (MOS) test [1]. Subjective testing is expensive, time-consuming, and not suitable for real-time applications such as online quality monitoring. Objective quality measurement methods are preferred in many applications. Objective methods can be classified as either link-parametric or signal based. Link-parametric methods use connection parameters to estimate the subjective MOS. For IP networks, parameters can include codec and PLC type, packet loss pattern (random or bursty), loss rate, jitter, and delay. Representative algorithms include VQmon by Telchemy [2], PSI by Psytechnics [3], and the International Telecommunications Union (ITU) E-model [4]. Pure link-parametric methods do not account for distortions that are not captured by the link parameters. Sources of such distortions include acoustic noise and tandem connections with analog links that do not convey upstream equipment and signal conditions downstream. The E-model is actually recommended for use as a transmission planning tool and is not recommended for quality measurement.

Signal based methods use perceptual features extracted from the speech signal to estimate quality. Signal based approaches can be classified as double- or single-ended, depending on whether a clean reference signal is required or not, respectively. ITU-T Recommendations P.862 [5] (PESQ) and P.563 [6] represent the current state-of-art double- and single-ended standard algorithms. Standard algorithms can achieve quality estimates that are highly correlated with subjective scores. Experiment results described herein, however, suggest that high estimation error is incurred for PLC-processed speech, particularly so for P.563. Thus, the applicability of the standard algorithms to VoIP is questionable.

In this paper, a hybrid signal-and-link-parametric approach to single-ended quality measurement of packet speech is proposed. The method makes use of IP connection parameters as well as the speech signal for quality estimation. Connection parameters such as codec and PLC type, packet size, and loss pattern are used to determine a "base quality" representative of a specific type of transmission links. Degradation factors are then computed from perceptual features extracted from the speech signal and are used to adjust the base quality accordingly. Degradation factors are based on Kullback-Leibler distances (KLD) between a Gaussian mixture (GM) model trained online for the extracted features and GM reference models of normative speech behavior. Experiments show that the hybrid approach reduces the root-mean-square estimation error ($RMSE$) in comparison with PESQ and P.563.

2. PLC IMPAIRMENT MODELING

It is known that perceived quality of packet loss concealed speech varies with codec and PLC types, packet size, loss pattern, and loss rate [7]. Our experiments suggest that PESQ and P.563 accuracy is also sensitive to such parameters. As an example, for speech processed by the G.711 PLC algorithm

(with 1%-10% random losses), PESQ attains a correlation (R) between subjective MOS and objective MOS of 0.835 for 10ms packet sizes, whereas $R = 0.876$ is attained for 20ms packets. For the same conditions, P.563 attains $R = 0.749$ and $R = 0.513$, respectively. Similar accuracy sensitivity is found for different PLC types and loss patterns. This sensitivity motivates our approach of modeling PLC impairments.

2.1. Gaussian Mixture Reference Models

We design reference models of PLC impairments. Models are designed for perceptually-relevant features (described in Section 3.1) extracted from *active* frames of loss-concealed speech. It is known that quality degradation is more pronounced if losses occur during active speech segments. In fact, if background noise is stationary and comfort noise insertion is in place, losses that occur during inactive periods of speech have far less impact on perceived quality. Modeling is accomplished by representing the probability distribution of extracted features with a Gaussian mixture (GM) density. A GM density is a weighted sum of M component densities $f(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x})$, where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{x})$ are K-variate Gaussian densities with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$.

As an embodiment of the proposed approach, we design GM models for three codec-cum-PLC types, three packet sizes (except for one codec where one packet size is used), and random and bursty losses. A complete description of the degradation conditions is given in Section 4.2. A total of 15 GM models are designed, 14 for PLC impairments (2 codecs $\times$ 3 packet sizes $\times$ 2 loss patterns + 1 codec $\times$ 1 packet size $\times$ 2 loss patterns) and one for high-quality, undistorted speech. The latter models the normative behavior of clean speech. Previous research [8] has shown that clean speech models are useful for measuring the quality of speech degraded by conditions unseen to the algorithm, e.g., noise-corrupted speech. The parameters for the 15 models are stored in a lookup table and are used in KLD calculation, as described in Section 3.

2.2. Base Quality Calculation

We use subjectively scored training data to calculate a "base MOS" for each PLC condition considered above. This base MOS reflects the average quality of a transmission link composed of a specific PLC algorithm, operating on a specific packet size, under a given loss pattern. As done with the reference model parameters, the calculated base MOSs are stored in a lookup table and are used by the MOS mapping module depicted in Fig. 1. As will be described in the sequel, the base MOS is adjusted according to degradation factors calculated from features extracted from the test speech signal.

3. ALGORITHM ARCHITECTURE

The overall architecture of the proposed algorithm is depicted within the dotted lines in Fig. 1. Connection parameters,

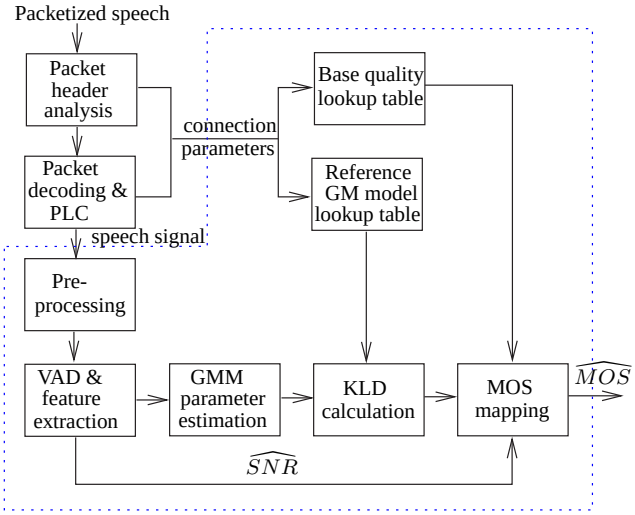

Fig. 1. Architecture of the proposed algorithm.

extracted from real-time protocol (RTP) or real-time control protocol (RTCP) headers, are used as input to the base MOS and the GM reference model lookup tables. Header-extracted parameters include codec-cum-PLC type and frame size. We assume that the loss pattern type is appropriately determined at the packet receiver/decoder; loss pattern type also serves as input to the lookup tables. Once packets are decoded and PLC is performed, the speech signal is level-normalized and filtered. Perceptual features are then extracted from the pre-processed test signal every 10ms. The voice activity detector (VAD) labels the feature vector of each frame as either active or inactive; only active vectors are kept. The features extracted from the test signal are represented by a GM density whose parameters are estimated using the expectation-maximization (EM) algorithm. Two KLDs are then calculated between the online-estimated GM model and two reference models. One KLD measures the similarity between the online model and the clean speech reference model; the other, between the online model and a link-specific PLC reference model. The calculated distances, together with a base MOS and an estimated signal-to-noise ratio ($\widehat{SNR}$) are mapped to a final estimated MOS ($\widehat{MOS}$). A more detailed description of each block is provided in the remainder of this section.

3.1. Pre-processing, VAD and Feature Extraction

The pre-processing module performs level normalization and intermediate reference system (IRS) filtering. The level of the speech signal is normalized to -26dBov and a modified IRS filter is applied to emulate the handsets used in listening tests. Voice activity detection (VAD) is employed to label speech frames as active or inactive. Here, the VAD from ITU-T G.729B [9] is used. As will be shown in Section 4.2, the SNR estimated by the VAD algorithm ($\widehat{SNR}$) is useful for quality measurement of speech corrupted by additive noise prior

to packetization. As for perceptually relevant features, we investigate the effectiveness of using p^{th} order perceptual linear prediction (PLP) [10] cepstral coefficients, $\mathbf{x} = \{x_i\}_{i=0}^{p}$. The zeroth cepstral coefficient is used as a log-energy term. We also experiment with delta and double-delta coefficients [11] as measures of signal spectral dynamics.

It is important to highlight that in some specific VoIP codec implementations, VAD decisions and spectral parameters are transmitted by the encoder and are readily available at the decoder. Such information can be used by the proposed algorithm to assist in VAD and feature extraction tasks. The architecture in Fig. 1 is for conceptual clarity and does not account for integrated processing that advantageously reduces overall computation complexity of the proposed scheme.

3.2. Online GMM Parameter Estimation

Online, the expectation-maximization (EM) algorithm is used to train a GM density on features extracted from the test signal. It is known that for diagonal covariance GMMs, the number of parameters that needs to be estimated during training is given by $M(2K+1)$, where $K = p+1$. Since our databases comprise speech files that have an average length of 7 seconds and an activity ratio of 70-90%, we restrict our experiments to 5^{th} order PLP coefficients ($p = 5$). Choosing $p = 5$ results, on average, in a training ratio (ratio between number of frames in the test signal and number of parameters estimated during training) that is greater than 10.

An alternate approach is taken when we investigate augmenting PLP coefficients with delta and double-delta coefficients. In this case, accurate GM modeling on a "per-sample" basis is not warranted as the training ratio is very low. A possible solution, and the one used here, is to perform GM modeling on a "per-condition" basis, where all files that belong to a specific degradation condition are used for parameter estimation. In analogy to quality monitoring of an ongoing call, per-condition accumulation can be viewed as the equivalent of collecting speech packets until a certain training ratio is obtained. An entailing condition is that loss statistics remain unchanged during collection. Our approach allows for encapsulating a much wider range of link conditions in the tables in Fig. 1 than our present concept-demonstration experiment.

3.3. KLD Calculation and MOS Mapping

The proposed algorithm makes use of the Kullback-Leibler distance to compute a degradation factor that is used to adjust the base MOS. The KLD [12] measures the dissimilarity between two probability density functions. Here, two KLDs are calculated. One measures how well the online-estimated model approximates the reference PLC model; the other, how well it approximates the clean speech model. We use a simplified version of the KLD described in [13]. In the following equations, f represents the reference model, $\tilde{f}$ the online estimated model, and $D(f, \tilde{f})$ describes how well $\tilde{f}$ approximates f. The simplified KLD is given by

$$D(f, \tilde{f}) = \sum_{i=1}^{M} \sum_{j=1}^{\tilde{M}} \alpha_i \tilde{\alpha}_j D(b_i(\mathbf{x}), \tilde{b}_j(\mathbf{x})) \tag{1}$$

where

$$D(b_i(\mathbf{x}), \tilde{b}_i(\mathbf{x})) = \frac{1}{2}\Big(\log\Big(\frac{\det \tilde{\mathbf{\Sigma}}_i}{\det \mathbf{\Sigma}_i}\Big) + \mathrm{trace}(\tilde{\mathbf{\Sigma}}_i^{-1}\mathbf{\Sigma}_i) + (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^T \tilde{\mathbf{\Sigma}}_i^{-1}(\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) - K\Big) \tag{2}$$

is the KLD between two K-variate Gaussian densities. The KLD is an asymmetric measure, i.e., $D(f, \tilde{f}) \neq D(\tilde{f}, f)$. We use the symmetrized measure $D_{sym}(f, \tilde{f}) = [D(f, \tilde{f})^{-1} + D(\tilde{f}, f)^{-1}]^{-1}$ [14] which we found to improve performance.

Lastly, the two KLDs, together with the base MOS and $\widehat{SNR}$, are mapped to an estimated MOS. As will be shown in Section 4.2, $\widehat{SNR}$ is a useful measure for cases where the speech signal is corrupted by additive noise prior to transmission through an IP network. We experiment with linear, multivariate polynomial and support vector regression (SVR) as candidate mapping functions. Simulation results showed that a radial basis SVR, with parameters optimized via linear search, provides least estimation error. The results below are all based on using SVR.

4. EXPERIMENTAL SETUP

4.1. Database Description

Two subjectively scored multilingual databases (English and French) are used in our experiments; subjective testing is performed conforming to [1]. The databases comprise speech processed by G.711, G.729 and Adaptive Multi-Rate (AMR) codecs. The PLC used for G.711 is described in [15]. G.729 and AMR have their own built-in PLCs. Packet sizes are 10, 20, and 30ms, except for AMR where only 20ms packets are used. Random and bursty losses are simulated at 1, 2, 4, 7, and 10%. Losses are applied to active packets only; a packet is classified as active if it comprises active speech frames only. Speech files corrupted by three noise types (hoth, babble, car) at two SNR levels (10 and 20dB) prior to packetization are also included. In this situation, random and bursty losses are simulated at 2% and 4%. A simple PLC scheme consisting of silence insertion is also used. The French database is used to train the PLC models and the MOS mapping function; the English database is regarded as unseen and is used for testing. Each database comprises 824 speech files, of which 216 are noise-corrupted. A total of 206 degradations conditions are included in each database.

4.2. Simulation Results

Initial experiments were performed using only PLP cepstral coefficients as features. On the test database, a "per-condition"

Table 1. Performance comparison of PESQ, P.563 and the proposed algorithm

Figure of Merit	PESQ	P563	PLP-SNR	Deltas	Deltas-SNR
R	0.919	0.673	0.918	0.907	0.922
$RMSE$	0.562	0.797	0.366	0.361	0.331
$\%R$	–	–	36.4	34.8	37.0
$\%RMSE$	–	–	54.1	54.7	58.5

$R = 0.831$ and $RMSE = 0.454$ between subjective MOS and objective MOS is attained. If noise-corrupted PLC files are excluded, we obtain $R = 0.879$ and $RMSE = 0.391$. To improve estimation performance for the noise-corrupted files, we use the SNR estimated by the VAD algorithm as an added feature for the MOS mapping function. Table 1 compares the performance of PESQ, P.563, and the proposed algorithm. The labels "PLP" and "Deltas" indicate performance figures for models using PLP cepstra and PLP combined with delta and double-delta parameters, respectively. The suffix "-SNR" indicates that $\widehat{SNR}$ is also used. We do *not* perform 3^{rd} order polynomial mapping in an attempt to investigate the true $RMSE$ produced by the algorithms. Rows labeled "$\%R$" and "$\%RMSE$" list percentage increase in R and percentage reduction in $RMSE$, relative to P.563, respectively.

As can be seen, improved performance is attained if $\widehat{SNR}$ is used as a complementary feature. The performance attained with Deltas is comparable to that of PLP-SNR; thus, the Deltas configuration is preferable when accurate SNR estimates are not available. In all cases, the proposed algorithm has correlation results comparable to that of PESQ, while offering the benefit of not requiring access to the clean speech signal. Substantial improvement in both R and $RMSE$ is obtained relative to P.563. Most important, a reduction in $RMSE$ of up to 58% can be attained with Deltas-SNR. Relative to PESQ, the obtained reduction is of 41%. These results suggest that the proposed algorithm is a better candidate for online monitoring of speech quality. Furthermore, it is important to emphasize that a 35% reduction in algorithm processing time is attained relative to P.563. The comparison is performed between a Matlab implementation of the proposed algorithm and the ANSI-C reference implementation of P.563; a complete C implementation of the proposed algorithm would further increase the speedup.

5. CONCLUSIONS

A hybrid signal-and-link-parametric quality measurement algorithm for packetized speech is proposed. The method improves on pure link-parametric approaches by measuring distortions that are not captured by connection parameters. Lower $RMSE$ is also attained relative to two standard signal based schemes; low $RMSE$ is a valuable attribute for online speech quality monitoring applications.

6. REFERENCES

[1] ITU-T P.800, "Methods for subjective determination of transmission quality," Intl. Telecom. Union, 1996.

[2] www.telchemy.com/vqmonep.html

[3] www.psytechnics.com/site/sections/products/psi.php

[4] ITU-T Rec. G.107, "The E-model, computational model for transmission planning," Intl. Telecom. Union, 2003.

[5] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Intl. Telecom. Union, 2001.

[6] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," Intl. Telecom. Union, 2004.

[7] L. Ding and R. Goubran, "Assessment of effects of packet loss on speech quality in VoIP," in *Proc. of the IEEE Int. Workshop on Haptic, Audio and Visual Environments and Their Applications*, Sep. 2003, pp. 49–54.

[8] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, March 2005, pp. 125–128.

[9] ITU-T Rec. G.729 - Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to Rec. V.70," Intl. Telecom. Union, 1996.

[10] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Society of America*, vol. 87, pp. 1738–1752, April 1990.

[11] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[12] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[13] Z. Liu and Q. Huang, "A new distance measure for probability distribution function of mixture type," in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 1, June 2000, pp. 616–619.

[14] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," Tech. Rep., 2001.

[15] ITU-T Rec. G.711-Annex I, "A high quality low-complexity algorithm for packet loss concealment with G.711," Intl. Telecom. Union, 1996.