Degradation-Classification Assisted Single-Ended Quality Measurement of Speech

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

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

Abstract

We propose an algorithm to classify speech degradations at network endpoints and to estimate the speech quality based on the degradation classification decision. Perceptual features from degraded speech signals are used to form statistical reference models of different degradation classes. Consistency measures, calculated between degraded speech signals and the reference models, are used to train a degradation classifier and mean opinion score (MOS) mappings. The quality of a received speech signal is estimated based on its degradation class and the MOS mapping associated with the class. Experimental results show that the proposed algorithm achieves high classification accuracy, and degradation classification improves the accuracy of the quality estimate.

Index Terms: speech communication network, speech degradations, speech transmission impairments, degradation classification, speech quality measurement.

1. Introduction

Quality assurance of speech communication networks becomes increasingly challenging with fast development of new speech communication technologies. While new technologies bring new services and lower costs, they also complicate end-to-end voice telephony connections and aggravate the uncertainty of voice quality. Today, voice transmitted over networks is exposed to a plethora of different sources of degradation, each with its own peculiar impairment to voice quality. Accurate identification of degradations enables deployment of appropriate corrective measures to assure the quality of service delivered. For real-time quality monitoring, knowing the degradation sources of an impaired speech signal helps to estimate the speech quality more accurately.

For end-to-end speech communication, acoustic noise is a significant factor of speech quality degradation. To combat ambient noise, noise suppression (NS) is introduced and has become an essential part of communications equipment, such as mobile and hands-free phones. While NS improves speech quality by reducing the background noise, it also distorts the speech signal and introduces annoying artifacts such as musical noise. Therefore, NS forms a distinctive type of degradation. Over the years we have seen a continuous migration of voice calls from conventional networks to Voice over Internet Protocol (VoIP) networks. In IP based telephony networks, packets can be lost and packet loss concealment (PLC) algorithms are used to recover from lost packets. While PLC algorithms partially recover the speech quality, the PLC enhanced speech signal still contains distortions due to packet loss and recovery, thus forming a specific type of degradation. Moreover, voice transmitted over heterogeneous networks may be processed by a sequence of codecs constituting a tandeming condition. Processing by a tandem of codecs may also result in noticeable degradations. Other sources of degradations may include circuit noise, bit errors, and echoes.

Accurate estimation of voice quality is essential for quality assurance of voice telephony networks. For this purpose, many objective quality measurement algorithms have been proposed. The ITU-T P.862 standard (PESQ), is the current "start-of-art" double-ended algorithm [1]. Single ended algorithms [2, 3, 4], on the other hand, do not depend on a reference signal therefore are more amenable for real-time speech quality monitoring. Most of these algorithms measure the received speech quality without analysis of the underlying speech degradation sources. Degradations from different sources have distinctive behaviors and certainly contribute differently to the impairment of speech quality. Therefore, speech quality might be more accurately estimated if degradations impairing the speech signal are known. Moreover, such knowledge facilitates the deployment of corrective measures, in real-time or otherwise [5].

In this paper, we describe an algorithm to classify speech degradations and to estimate speech quality based on the classification decision. Perceptual features of degraded speech are used to build separate degradation reference models, one for each degradation class. A clean speech reference model is also created. Given a degraded speech signal, consistency measures are computed for the perceptual features of the signal relative to each reference model. A degradation classifier maps the consistency measures to a classification decision. The speech quality is estimated based on the classification decision and a MOS mapping. The proposed algorithm achieves high classification accuracy and superior quality estimation performance.

2. Algorithm Description

2.1. Algorithm Overview

The architecture of the proposed algorithm is depicted in Fig. 1. Offline, perceptual features extracted from

degraded speech signals are used to train a reference model and a MOS mapping for each individual degradation class. A clean speech reference model is also trained. Each reference model comprises multiple Gaussian mixture models (GMMs). Consistency measures calculated with respect to the reference GMMs are used to train a speech degradation classifier and each degradation-specific MOS mapping.

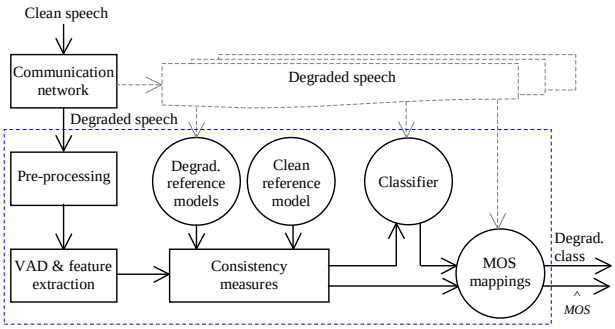

Figure 1: Architecture of the proposed algorithm.

Online, the received speech signal experiences transmission distortions in the communications network. In general, the signal can be subject to several types of degradations. As an initial step towards degradation classification, we assume in this paper that the signal is subject to only one type of distortion. In practice, the quality degradation may be mainly caused by one type of distortion, and/or the network may choose to ameliorate only the dominant distortion. The degraded speech signal is first pre-processed by the proposed algorithm, where the level of the speech signal is normalized. Perceptual features are then extracted from the preprocessed speech signal every 10 milliseconds. The voice activity detector (VAD) and voicing classifier label the feature vector of each frame as either voiced, unvoiced, or inactive. Consistency measures are calculated between the degraded speech signal and the GMM reference models. A degradation decision is then made by the degradation classifier. Once a degradation class is decided, a MOS mapping corresponding to the class is applied to obtain a MOS estimate for the degraded speech signal. A detailed description of each processing block of the algorithm is provided in the remainder of this section.

2.2. Pre-processing, VAD and Feature Extraction

The pre-processing module normalizes the speech signal level to -26 dBov. The VAD from ITU-T G.729B [6] and the voicing decision algorithm described in [7] are used to label the frames. As for perceptually relevant features, we investigate using p^{th} order perceptual linear prediction (PLP) [8] cepstral coefficients, $\mathbf{x} = \{x_i\}_{i=0}^{p}$. PLP cepstra exploit three essential psychoacoustic precepts (critical band spectral resolution, equal-loudness curve, and intensity loudness power law) and have been proven to be more consistent with human auditory perception than speech-production-based linear predictive analysis. Because of these attractive properties, PLP features have been used in speech quality estimation [4] and speech recognition [8], and are used here as features for degradation classification. We experiment with several PLP orders and $p = 5$ is chosen as it strikes a balance between performance and complexity.

2.3. Reference GMMs and Parameter Estimation

Reference modeling of each degradation class is accomplished by using GMMs to represent the probability distribution of PLP features from the class. A GM density is a weighted sum of M component densities $p(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x})$, where $\alpha_i \geq 0, i = 1, \cdots, M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{x})$ are K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. $\boldsymbol{\lambda}$ represents the GMM parameter set $\{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{M}$, which is estimated by an EM algorithm.

Reference models are created for four degradation classes: acoustic noise, noise suppression, codec-tandeming, and packet loss concealment. A clean speech class is also formed to distinguish high-quality speech signals from degraded speech signals. For each degradation class, and the clean speech class, three GMMs are designed, one for each of the three different speech frame types (voiced, unvoiced, inactive). Hence, each GMM can be denoted by $p_{\text{class,frame}}(\mathbf{x}|\boldsymbol{\lambda})$, where the subscript "class" represents a degradation class or the clean speech class, and the subscript "frame" indicates a frame type.

2.4. Consistency and Classifier Design

In principle, by evaluating the density of a reference GMM $p_{\text{class,frame}}(\mathbf{x}|\boldsymbol{\lambda})$ using a feature vector $\mathbf{x}$ from the received signal, a measure of consistency between the feature vector and the reference GMM is obtained. Thus, the consistency between an observed speech signal and a reference GMM is defined as the normalized log-likelihood

$$c_{\text{class,frame}}(\mathbf{X}) = \frac{1}{N_{\text{frame}}} \sum_{j=1}^{N_{\text{frame}}} \log(p_{\text{class,frame}}(\mathbf{x}_j|\boldsymbol{\lambda})), \quad (1)$$

where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N_{\text{frame}}}$ are the feature vectors that belong to a given frame type, and N_{frame} represents the total number of such feature vectors. Normalization is required as N_{frame} varies for different test signals. Larger $c_{\text{class,frame}}$ indicates greater consistency. In the rare case of $N_{\text{frame}} = 0$, a constant $c_{\text{class,frame}} = c = -10$ is assigned. For each test signal, a total of fifteen consistency measures are computed, one for each of the five classes and three frame types.

Online, the consistency measures of the test speech signal serve as input to a trained degradation classifier for a degradation class decision. Offline, consistency measures of degraded speech signals, along with their respective degradation class information, are used to train the degradation classifier. We experiment with two candidate classifiers: Support Vector Machine (SVM) and Classification and Regression Tree (CART) [9]. Simulation results indicate that the SVM based classifier achieves lower classification error rate. The classification results presented in Section 3.2 are all based on using SVM.

2.5. MOS Mapping Design

The consistency measures are also used to train each degradation-specific MOS mapping. In a previous work

[10], consistency measures are calculated with respect to a 'global' degradation reference model and a clean speech reference model, and there is only one MOS mapping. With degradation classification, reference models and MOS mappings can be tailored to each degradation class. It is possible for the class dependent MOS mapping to provide a more accurate MOS estimate than the global MOS mapping, provided that the degradation class is correctly determined by the classifier. Hence, it is of interest to compare the quality estimation performance of using the classifier against the performance of using the global model, as will be done in Section 3.3.

The design of MOS mappings using consistency measures is as follows. Offline, a weighted consistency measure is formed by taking the product of the consistency measure (1) and the fraction of the frame type in the speech signal. The weighted consistency measures of degraded speech signals, along with their respective subjective MOS data, are used to train the MOS mappings. Each mapping is optimized for a given degradation class and uses consistencies calculated with respect to the clean speech model and the specific degradation model. Online, the MOS mapping that corresponds to the classification decision is applied to produce a MOS estimate for the test signal. We experiment with Support Vector Regressor (SVR) and Multivariate Adaptive Regression Splines (MARS) [9] as MOS mapping functions. Simulation results indicate that SVR attains better performance. The quality estimation results presented in Section 3.3 are all based on using SVR.

3. Experimental Results

3.1. Database Description

The degraded and clean speech signals used in our experiments are taken from two publicly available databases (NOIZEUS [11] and ITU-T P-Series Supp. 23 [12]) and one proprietary database. The NOIZEUS database contains speech signals corrupted by eight different types of real-world noise at four SNR levels (0, 5, 10, and 15 dB), and is used to represent degradations due to acoustic noise. Furthermore, each noise corrupted speech signal is processed by thirteen different NS algorithms. The NS algorithms can be divided into four categories: spectral subtractive, subspace, statistical-model, and Wiener [11].

The proprietary database contains speech signals processed by G.711, G.729 and Adaptive Multi-Rate (AMR) codecs, with packet loss concealment (PLC). Random and bursty losses are simulated at 4%, 7%, and 10%, respectively. These speech signals are used to represent the degradation class of PLC. Moreover, the database also contains speech signals corrupted by four types of noise (white, car, street, babble) at three SNR levels (0, 10, 20dB). These speech signals are used as additional acoustic noise degradation conditions. The ITU-T Supp. P. 23 Experiment 1 database has a variety of codec tandeming conditions involving ITU-T speech coding standards (G.729, G.726, and G.728) and codecs (Full-Rate GSM, IS-54 and Half-Rate JDC) deployed in digital mobile radio systems. Lastly, clean speech signals are selected from the ITU-T Supp. P. 23 Experiment 1 and 3 databases. The aforementioned speech databases are organized by the degradation conditions described above and each of the four degradation classes under study is represented by a range of such conditions.

Table 1: Confusion matrix for classification test set

Actual Category	Predicted Category					Accuracy Rate (%)
	Acoustic	NS	PLC	Tandem	Clean	
Acoustic	470	10	0	0	0	97.9
NS	31	797	2	2	0	95.8
PLC	1	0	167	0	0	99.4
Tandem	0	0	0	144	0	100
Clean	0	0	0	1	663	99.8
Average	–	–	–	–	–	**97.9**

Speech signals from each degradation class are divided into three groups. The first group, called the training dataset, is used to train the GMM reference models. The second group, called the validation dataset, is used to train the MOS mapping for each degradation class and to train the degradation classifier. The third group, called the test dataset, comprises speech signals regarded as unseen and is used for classification and quality estimation testing. For quality estimation, the test dataset consists of 724 speech signals, with 63 degradation conditions. Since subjective MOS data is not required for degradation classification, the test dateset for testing the classifier comprises 2288 speech signals.

3.2. Classification Performance

Table 1 presents the classification result in terms of a confusion matrix. The five test classes are denoted by "Acoustic," "NS," "PLC," "Tandem," and "Clean," respectively in the table. Each element of the confusion matrix represents the number of test signals that are classified to the predicted category. Therefore, elements on the diagonal represent the number of correctly classified signals, while the others represent the number of misclassified signals. Classification accuracy rate, for each degradation class, is reported on the last column on the right. The average accuracy rate is also reported.

As can be seen, the proposed algorithm achieves high classification accuracy for most degradation classes, except for some classification errors between acoustic noise and noise suppression. An average accuracy rate of 97.9% is attained. Further investigation reveals that the 31 NS signals misclassified as "Acoustic" contain a fair amount of background noise after suppression. Moreover, of these misclassified signals, 30 are from three NS algorithms (two spectral subtractive and one Wiener based) which seem to provide insufficient noise reduction. This experiment also tests the robustness of the proposed algorithm. Test signals are comprised of degradation conditions unseen to the classifier, such as packet loss at different rates and different types of acoustic noise, as well as languages other than the ones used in training; classification performance is, therefore, promising.

Table 2: Confusion matrix for quality estimation test set

Actual Category	Predicted Category					Accuracy Rate (%)
	Acoustic	NS	PLC	Tandem	Clean	
Acoustic	203	1	0	0	0	99.5
NS	8	199	1	0	0	95.7
PLC	1	0	167	0	0	99.4
Tandem	0	0	0	144	0	100
Average	–	–	–	–	–	**98.5**

Table 3: Performance comparison between the global model based algorithm and the classifier based algorithm

	Global Model Based		Classifier Based			
	R	ϵ	R	$\%R$	ϵ	$\%\epsilon$
Per Cond.	0.753	0.374	0.831	10.4	0.308	17.6
Poly. Regress.	0.772	0.341	0.851	10.2	0.282	17.3

3.3. Single-Ended Quality Estimation Performance

Here, we compare the performance of the proposed single-ended quality measurement algorithm and the approach proposed in [10], where a single global MOS mapping is devised. Due to degradation-specific MOS mappings, the performance of the proposed algorithm is directly impacted by the performance of the classifier. It is, thus, important to also test the effect of degradation misclassification on quality estimation performance. Table 2 shows the confusion matrix for the quality estimation test set. As can be seen, classification performance similar to that presented in Table 1 is attained.

The results in Table 3 serve to compare the proposed algorithm with the "global degradation model" paradigm of [10]. Per-condition correlation (R) and root-mean-square error (ϵ) between subjective MOS and estimated MOS are presented, for the two compared algorithms. The per-condition results after 3^{rd} order monotonic polynomial regression, as recommended in [2], are presented as well. The table also shows performance percentage improvement attained by the proposed algorithm over the global model based algorithm, in the two columns labelled "$\%R$" and "$\%\epsilon$". It can be seen that the proposed algorithm outperforms the global model based algorithm, indicating that quality estimation performance can be improved by using degradation classification. A 10.2% increase in R and 17.3% decrease in ϵ can be attained on the test dataset. For comparison, P.563 [2] achieves $R = 0.668$ and $\epsilon = 0.398$ on the same test dataset, after 3^{rd} order monotonic polynomial regression. P.563 attains relatively poor performance on NS and PLC enhanced speech, corroborating prior results [4] and [13].

The performance figures of the proposed algorithm in Table 3 correspond to the classification results shown in Table 2. In the ideal case where all test signals are correctly classified, the proposed algorithm attains best possible performance: $R = 0.854, \epsilon = 0.279$ after 3^{rd} order monotonic polynomial regression. Thus, the performance loss due to misclassification is small as long as the classifier maintains a high accuracy rate.

4. Conclusions

An algorithm is proposed for speech degradation classification and applied to improve quality estimation performance. The classifier is shown to achieve accurate performance on an unseen test dataset. When used to assist in a single-ended speech quality measurement task, an improvement in R and in RMSE of approximately 10.2% and 17.3%, respectively, is attained.

5. References

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

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

[3] D.-S. Kim and A. Tarraf, "Enhanced perceptual model for non-intrusive speech quality assessment," in *Proc. Int. Conf. Acous., Speech, Sig. Proc.*, vol. I, May 2006, pp. 829–832.

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

[5] S. Mohamed, F. Cervantes-Pérez, and H. Afifi, "Integrating network measurements and speech quality subjective scores for control purposes," in *INFOCOM*, vol. 2, 2001, pp. 641–649.

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

[7] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT)*. Elsevier Science Publishers, pp. 495–518, in Speech Coding and Synthesis, 1995.

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

[9] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[10] T. H. Falk and W.-Y. Chan, "Enhanced non-intrusive speech quality measurement using degradation models," in *Proc. Int. Conf. Acous., Speech, Sig. Proc.*, vol. 1, May 2006, pp. 837–840.

[11] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. Int. Conf. Acous., Speech, Sig. Proc.*, May 2006, pp. 153–156.

[12] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," Intl. Telecom. Union, Feb. 1998.

[13] T. H. Falk, H. Yuan, and W.-Y. Chan, "A hybrid signal-and-link-parametric approach to single-ended quality measurement of packetized speech," in *Proc. Int. Conf. Acous., Speech, Sig. Proc.*, vol. 4, April 2007, pp. 841–844.