Classification of Speech Degradations at Network Endpoints Using Psychoacoustic Features

Hua Yuan, Tiago H. Falk and Wai-Yip Chan
Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
E-mail: {yuanh, falkt, chan}@ee.queensu.ca

***Abstract*—We propose a method of classifying speech degradations at network endpoints. Perceptual features are extracted from degraded speech signals and used to form statistical reference models of behaviors of different degradation types. Consistency values between degraded speech signals and the reference models are calculated and used to train a degradation classifier. The consistency values of a received degraded speech signal then serve as predictors in the trained classifier for a degradation type decision. The proposed method is tested on four commonly encountered degradation types with unseen data and the experimental results show that the method achieves high classification accuracy. The proposed method can be used to enhance applications such as speech enhancement, recognition, and quality estimation.**

Keywords: speech impairment, speech quality, telephony network, degradation classification, perceptual features, Gaussian mixture model, consistency measure, support vector machine.

I. INTRODUCTION

The number and types of networks that can mediate a voice telephony connection continue to grow at a fast pace. Over years we have seen a developing network shift from analog to digital, wired to wireless, and a continuous migration of some voice calls from conventional networks to Voice over Internet Protocol (VoIP). While this development brings new services and lowers costs, it also exacerbates the uncertainty of end-to-end voice quality. Voice transmitted over multi-stage, hybrid networks nowadays is impaired by more kinds of degradations than ever. In the IP based telephony network, packets can be lost due to network delay, congestion or errors, thus causing degradation in voice quality. Packet loss concealment (PLC) algorithms are used to recover from lost packets and to improve the impaired quality. Voice transmitted over heterogeneous networks may be processed by a sequence of codecs that constitutes a tandeming condition. While a single codec can degrade speech quality noticeably by coding at low bit-rates (LBR), as found in some wireless codecs, processing by a tandem of codecs, whether LBR codecs are used or not, may also result in noticeable degradations. Other sources of degradations may include background acoustic noise, circuit noise, bit errors, and echoes.

For dynamic quality assurance, real-time voice degradation monitoring and call control can be deployed. Accurate identification of the degradations impairing a received speech signal enables initiation of appropriate corrective measures. Impaired speech signals can be more efficiently enhanced by identifying the sources of degradations than treating them identically. Besides, for dynamic quality assurance, a real-time speech quality estimator is desired. Most online speech quality estimation algorithms, including the current state-of-the-art standard ITU-T P.563 [1], measure the received speech quality without knowing the sources of degradations. Degradations from different sources have distinctive behaviors and certainly contribute differently to the decrease of speech quality. Therefore, the performance of current speech quality estimation algorithms can be improved if the degradation type information is available.

In this paper, we describe an algorithm to classify speech degradations at network endpoints. The algorithm uses perceptual features extracted from degraded speech signals to form reference models of behaviors of different degradation types. Modeling is accomplished by representing the probability distribution of the perceptual features with a Gaussian mixture (GM) density. Classification is achieved by means of a consistency measure, calculated between the degraded test speech signal and the reference models. Here, two classification schemes are investigated. Four commonly encountered degradation types are tested and both schemes are shown to achieve high classification accuracy on an unseen test dataset.

II. ALGORITHM DESCRIPTION

Fig. 1 shows the architecture of the proposed algorithm. During transmission over the communication network, the speech signal experiences various kinds of degradations. The received degraded speech signal at

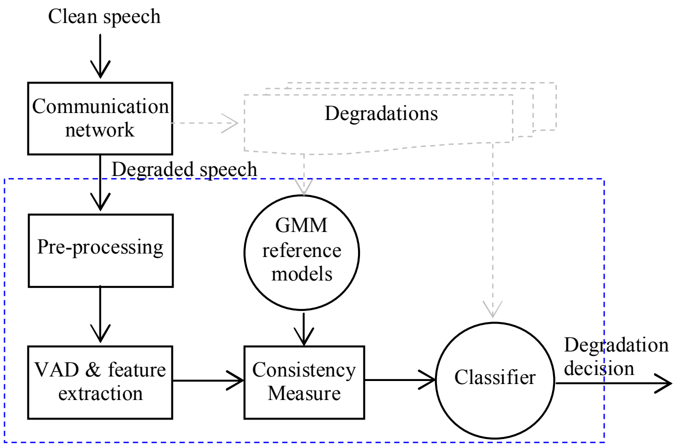

Figure 1. Architecture of the proposed algorithm.

network endpoints is first pre-processed by the algorithm, where the level of the speech signal is normalized. Perceptual features are then extracted from degraded speech signal every 10 milliseconds. The voice activity detector (VAD) labels the feature vector of each frame as either active or inactive.

Offline, degraded speech signals from training datasets are used to produce reference models for each degradation type and to train a classifier. Perceptual features are extracted. The probability distribution of the features is modeled with a Gaussian mixture model (GMM) and the model parameters are estimated by an expectation-maximization (EM) algorithm. Consistency measures with degradation reference models are then calculated for each degraded speech signal and the consistency values are used to train a degradation classifier.

During online operation, the feature vectors of the received degraded speech signal are fed into the GMM reference models for consistency calculation. The classifier is then applied to produce a degradation decision. A detailed description of each processing block of the algorithm is provided in the remainder of this section.

A. Pre-processing, VAD and Feature Extraction

The pre-processing module performs level normalization so that the speech signal level is normalized to -26dBov. Voice activity detection (VAD) is employed to label speech frames as either active or inactive. Here, the VAD from ITU-T G.729B [2] is used.

As for perceptually relevant features, we investigate the effectiveness of using p^{th} order perceptual linear prediction (PLP) [3] 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 are used in applications of speech quality estimation [4] and speech recognition [5], and also used here as candidate features for degradation classification. We experiment with several PLP orders and $p = 5$ is chosen as it strikes a balance between performance and complexity. PLP features are extracted from the speech signal every 10 milliseconds and stored in a feature vector per frame.

B. GM Reference Models and Parameter Estimation

The reference model of each degradation type is created using PLP features extracted from active frames of degraded speech signals. As an initial approach of the problem, we focus on modeling the active speech segments because the degradations considered here are perceptually more prominent in the active segments. Modeling is accomplished by using GMMs to represent the probability distribution of PLP features. 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 complete GMM parameter set $\{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$. As an embodiment of the proposed method, we design reference models for the following degradation types: acoustic noise, codec-tandeming, coding at low bit rates, and packet loss concealment. We also design a clean speech reference model to distinguish high-quality speech signals from degraded speech signals. Therefore, altogether five GMM reference models $p_{class \in \mathbf{S}}(\mathbf{x}\,|\,\boldsymbol{\lambda}), \mathbf{S} = \{1,2,3,4,5\}$, are created. Subscript “class” represents the index of the four degradation types and the clean speech type. We experiment with GMMs with 16 Gaussian components $(M = 16)$ and diagonal covariance matrices.

C. Consistency Measure and Degradation Classification

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

$$C_{class}(\mathbf{x_1},\ldots,\mathbf{x_N}) = \frac{1}{N}\sum_{j=1}^{N}\log(p_{class}(\mathbf{x_j} \mid \boldsymbol{\lambda})), \qquad (1)$$

where $\mathbf{x_1},\ldots,\mathbf{x_N}$ are the feature vectors of the speech signal, and N represents the total number of feature vectors. Normalization is required as N varies for different test signals. For each test signal, a total of five consistency measures are computed: one for each reference model, as distinguished by the subscript “class”. Since Larger C_{class} indicates greater consistency, a simple degradation decision can be made by picking the degradation type with the largest consistency value:

$$\text{Decision} = \arg\max_{class \in \mathrm{S}} C_{class}(\mathbf{x_1},\ldots,\mathbf{x_N}). \qquad (2)$$

A more sophisticated decision algorithm is obtained by applying a classifier. Offline, the consistency measures of degraded speech signals from training datasets, along with their respective degradation class information, are used to train a classifier. Online, the consistency measures of the test speech signal serve as input to the trained classifier for a degradation decision. We experiment with two candidate classifiers that are based on Support Vector Machine (SVM) [6] and Classification and Regression Tree (CART) [7], respectively. Simulation results indicate that the SVM based classifier achieves a lower classification error rate.

III. Experimental results

A. *Database Description*

The degraded and clean speech signals used in our experiments are taken from two publicly available databases (NOIZEUS [8] and ITU-T P-Series Supplementary 23 [9]) and two proprietary databases (LBR Codec and VoIP). The NOIZEUS database is comprised of speech signals corrupted by eight different types of real-world noise at four different SNR levels (0, 5, 10, and 15 dB), and hence is used to represent degradation due to acoustic noise. Speech files selected from the LBR Codec database are coded by seven vocoder algorithms (FS-1016, MBE-2.4, MBE-4.8, LPC-10E, STC-2.4A, STC-2.4B and STC-4.8) at two low bit-rates (2.4 and 4.8 Kbps), representing degradation produced by low bit-rate codecs. The VoIP database contains speech files 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%. 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. Finally, clean speech files are selected from the ITU-T Supp. P. 23 Experiment 1 and 3 databases, where speeches are spoken in different languages by different speakers.

TABLE I. NUMBER OF TEST FILES FOR EACH DEGRADATION TYPE

	Degradation Type				
	Acoustic noise	***LBR***	***PLC***	***Tandem***	***Clean speech***
Num. of Files	360	285	168	144	224

Speech files from the above mentioned databases are divided into two groups. One group is used to train the GMM reference models and the degradation classifier, while the other is regarded as unseen and used for testing. The total number of test files for each degradation type is given in Table I. In an attempt to test the robustness of the proposed algorithm to unseen degradation conditions, the signals in the test set are corrupted by degradation conditions unseen to training. Languages, other than the ones used in training, are also available in the test set in order to investigate the effectiveness of the proposed algorithm to unseen languages.

B. *Simulation Results*

As mentioned previously, we examine five degradation types for classification as denoted by “Acoustic noise”, “LBR (low bit-rate codecs)”, “PLC (packet loss concealment)”, “Tandem (codec tandeming)”, “Clean speech” in Table I. We experiment with two different classification methods: the one described by (2) and the SVM approach described in Section II. Table II and III present classification results as confusion matrices. Each element in the matrices indicates the number of test signals that are classified to the predicted category. Therefore, elements on the diagonal represent the total number of correctly classified,

TABLE II. CONFUSION MATRIX OF CLASSIFICATION DECISIONS BASED ON SIMPLE CONSISTENCY MEASURE

Actual Category	**Predicted Category**					**Accuracy Rate (%)**
	Acoustic noise	***LBR***	***PLC***	***Tandem***	***Clean speech***	
Acoustic noise	360	0	0	0	0	100
LBR	0	261	14	0	10	91.6
PLC	0	0	168	0	0	100
Tandem	0	0	7	137	0	95.1
Clean speech	0	18	0	9	197	88.0
Average	-	-	-	-	-	95.1

TABLE III. CONFUSION MATRIX OF CLASSIFICATION DECISIONS BASED ON TRAINED SVM CLASSIFIER

Actual Category	**Predicted Category**					**Accuracy Rate (%)**
	Acoustic noise	***LBR***	***PLC***	***Tandem***	***Clean speech***	
Acoustic noise	360	0	0	0	0	100
LBR	0	284	0	0	1	99.6
PLC	0	0	168	0	0	100
Tandem	0	0	8	134	2	93.1
Clean speech	0	2	0	0	222	99.1
Average	-	-	-	-	-	98.9

while the others represent the number of misclassified files. A classification accuracy rate is reported on the last right column of the tables for each degradation type. The tables also show an average accuracy rate, which is the weighted sum of accuracy rates of all degradation types, at the bottom, where the weight is given by the file number percentage of each degradation type.

As can be seen, the proposed algorithm achieves high classification accuracy, for both the consistency measure based and the SVM classifier based schemes. When compared with one the other, the SVM based scheme improves performance for two degradation types: LBR and Clean speech, thereby attaining a higher average accuracy rate.

The consistency measure based scheme, on the other hand, has an advantage of being simple. No classifier needs to be trained and the decision can be made directly upon the degradation type with the largest consistency value. The hindrance is that it attains a relatively low accuracy rate of 88% for clean speech detection. In fact, by applying a state-of-the-art speech quality estimation algorithm such as ITU-T P.563 [1] to the online received speech signal before performing degradation classification, the 'cleanness' of the speech signal can be pre-determined, as depicted in Fig. 2. Clean speech signals are detected if an objective quality score above a certain threshold is attained. If a 5-point mean opinion score (MOS) scale [4] is used, our experiments suggest that the threshold can be set to 3.8. In such scenario, overall classification accuracy of the simplified scheme increases to 97%.

It is observed that the proposed algorithm based on the trained SVM classifier attains high classification accuracy in nearly every degradation category except codec tandeming. Considering the test datasets comprise speech signals with many conditions unseen to the training process, the proposed algorithm is very robust and effective.

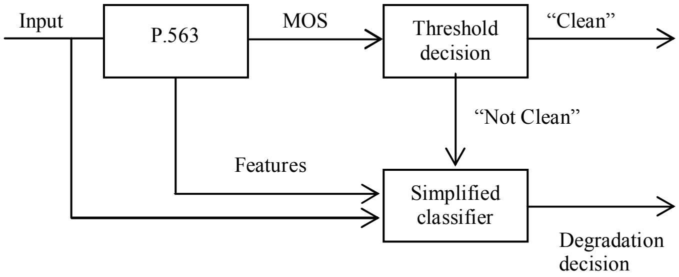

Figure 2. Architecture of simplified classifier working with P.563 .

IV. CONCLUSIONS

A method of classifying speech degradations at network endpoints is proposed based on statistical modeling of perceptual features of degradation types and a trained classifier using consistency measures between the degraded speech signals and the reference models. Experimental results show that the proposed method achieves promising degradation classification performance.

REFERENCES

[1] ITU-T P.563, Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications, Int. Telecommun. Union, Geneva, Switzerland, May 2004.

[2] 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.

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

[4] T. H. Falk and W.-Y. Chan, "Single-Ended Speech Quality Measurement Using Machine Learning Methods," *IEEE Transactions on Audio, Speech and Language Proc.,* Special Issue on Objective Quality Assessment of Speech and Audio, Nov. 2006.

[5] B. Kingsbury and N. Morgan, "Recognizing Reverberant Speech with RASTA-PLP," *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*-Volume 2, pp.1259, April 21-24, 1997.

[6] Vladimir Vapnik, Statistical Learning Theory, Wiley, NY, 1998.

[7] Breiman L, Friedman JH, Olshen RA, and Stone CJ, Classification and Regression Trees, Chapman & Hall (Wadsworth, Inc.): New York, 1984.

[8] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," *Proceedings of ICASSP-2006*, I, pp. 153-156, Toulouse, France, May 2006.

[9] "ITU-T coded-speech database," ITU-T, P-series Supplement 23, 1998.