EURASIP Journal on Applied Signal Processing 2005:9, 1410–1424

Objective Speech Quality Measurement Using Statistical Data Mining

Wei Zha

Power, Acquisition and Telemetry Group, Schlumberger Technology Corporation, 150 Gillingham Lane, MD 1, Sugar Land, TX 77478, USA
Email: wzha@ee.queensu.ca

Wai-Yip Chan

Department of Electrical & Computer Engineering, Queen's University, Kingston, ON, Canada K7L 3N6
Email: chan@ee.queensu.ca

Received 7 November 2003; Revised 3 September 2004

Measuring speech quality by machines overcomes two major drawbacks of subjective listening tests, their low speed and high cost. Real-time, accurate, and economical objective measurement of speech quality opens up a wide range of applications that cannot be supported with subjective listening tests. In this paper, we propose a statistical data mining approach to design objective speech quality measurement algorithms. A large pool of perceptual distortion features is extracted from the speech signal. We examine using classification and regression trees (CART) and multivariate adaptive regression splines (MARS), separately and jointly, to select the most salient features from the pool, and to construct good estimators of subjective listening quality based on the selected features. We show designs that use perceptually significant features and outperform the state-of-the-art objective measurement algorithm. The designed algorithms are computationally simple, making them suitable for real-time implementation. The proposed design method is scalable with the amount of learning data; thus, performance can be improved with more offline or online training.

Keywords and phrases: speech quality, speech perception, mean opinion scores, data mining, classification trees, regression.

1. INTRODUCTION

"Plain old telephone service," as traditionally provided using dedicated circuit-switched networks, is reliable and economical. A contemporary challenge is to provide high-quality, reliable, and low-cost voice telephone services over nondedicated and heterogeneous networks. Good voice quality is a key factor in garnering customer satisfaction. In a dynamic network, voice quality can be maintained through a combination of measures: design planning, online quality monitoring, and call control. Underlying these measures is the need to measure user opinion of voice quality. Traditionally, user opinion is measured offline using subjective listening tests. Such tests are slow and costly. In contrast, machine computation ("objective measurement"), which involves no human subjects, provides a rapid and economical means to estimate user opinion. Objective measurement enables network service providers to rapidly provision new network connectivity and voice services. Online objective measurement is the only viable means of measuring voice quality, for the purpose of real-time call monitoring and control, on a network-wide scale. Other applications of voice quality measurement include evaluation of disordered speech [1] and synthesized speech [2].

Algorithms for objective measurement of speech quality can be divided into two types: single-ended and double-ended (see Figure 1). Double-ended algorithms need to input both the original ("clean") and degraded speech signals, whereas single-ended algorithms need only to input the degraded speech signal. Single-ended algorithms can be used for "passive" monitoring, that is, nonintrusively tapping into a voice connection. Double-ended algorithms are sometimes called "intrusive" because a voice signal known to the algorithm has to be injected into the transmit end. Nevertheless, Conway in [3] proposes a method that employs double-ended algorithms without intruding on an ongoing call. The method is based on measuring packet degradations at the receive end. The measured degradations are applied to a *typical* speech signal to produce a degraded signal. A double-ended algorithm is used to map the speech signal and degraded signal to speech quality.

The performance of objective measurement algorithms is primarily characterized by the accuracy of the user opinion scores produced by the algorithm, using the opinion scores

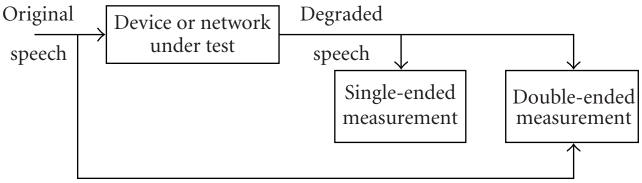


FIGURE 1: Single-ended and double-ended speech quality measurements.

obtained from subjective tests as accuracy benchmarks. The *mean opinion score* (MOS) [4], obtained by averaging the absolute categorical ratings (ACRs) produced by a group of listeners, is the most commonly used measure of user opinion. Subjective listening tests are generally performed with a limited number of listeners, so that the MOS varies with the listener sample and its size. In such a case, the degree of accuracy of objective scores can be assessed up to the degree of accuracy of the subjective scores used as benchmarks.

The International Telecommunications Union (ITU) standard [5, P.862], also called Perceptual Evaluation of Speech Quality (PESQ), is a double-ended algorithm that exemplifies the "state-of-the-art." An ITU standard for single-ended quality measurement [6, P.563] has recently reached a "prepublished" status. Objective measurement has the advantage of being consistent. While subjective tests can be used to estimate the MOS very accurately by using a large listening panel, objective measurement can provide a more accurate MOS estimate than a small listener panel (see [7] for a simple model of measurement variance). Hence, objective measurement, which can be automated and performed in real time, provides a very attractive alternative to subjective tests.

The process of human judgment of speech quality can be modeled in two parts. The first part, auditory perception, entails transduction of the received speech acoustic signal into auditory nerve excitations. Auditory models are well studied in the literature [8] and have been applied to the design of PESQ and other objective measurement algorithms [9, 10, 11]. Essential elements of auditory processing include bark-scale frequency warping and spectral power to subjective loudness conversion. The second part of the human judgment process entails cognitive processing in the brain, where compact features related to normative and anomalous behaviors in speech are extracted from auditory excitations and integrated to form a final impression of the perceived speech signal quality. Cognitive models of speech distortions are less well developed. Nevertheless, for the goal of accurate prediction of subjective opinion of speech quality, anthropomorphic modeling of cognitive processing is not strictly necessary.

In place of cognitive modeling, we pursue a statistical data mining approach to design novel double-ended algorithms. The success of statistical techniques in advancing speech recognition performance lends promise to the approach. Our algorithms are designed based on classifying perceptual distortions under a variety of contexts. A large pool of context-dependent feature measurements is created. Statistical data mining tools are used to find good features in the pool. Features are selected to produce the best estimator of the subjective MOS value. The algorithms demonstrate significant performance improvement over PESQ, at a comparable computational complexity. In effect, the statistical classifier-estimators serve as utilitarian models of human-cognitive judgment of speech quality.

This paper is organized as follows. Section 2 provides the background by introducing existing double-ended speech quality measurement schemes and two statistical data mining algorithms. Section 3 describes our speech quality measurement algorithm architecture, its basic elements and design framework, and feature design and mining. Lastly, in Section 4, various design methods and designed algorithms are examined and their performance are assessed experimentally.

2. BACKGROUND

In this section, we review briefly existing objective speech quality measurement methods and the statistical data mining techniques we have used.

2.1. Current objective methods

Early speech quality measures were used for assessing the quality of waveform speech coders. These measures calculate the difference between the waveform of the nondegraded speech and that of the degraded speech, in effect using waveform matching as a criterion of quality. Representative measures include the signal-to-noise ratio (SNR) and segmental SNR [12]. Measures of distortions in the short-time spectral envelopes of speech [13] were later introduced. These measures do not require the waveforms to match in order to produce zero distortion. They are suitable for low-bit-rate speech coders that may not preserve the original speech waveform, for example, linear-prediction-based analysis-by-synthesis (LPAS) coders [14]. For a comprehensive review of objective methods known till late 1980s, the reader can consult [15].

Measurement algorithms that exploit the human auditory perception rather than just the acoustic features of speech provide more accurate prediction of subjective quality. Representative algorithms include BSD (Bark spectral distortion) [11], MNB (measuring normalizing block) [9, 10], and PESQ (perceptual evaluation of speech quality) [5, 16]. A major difference among algorithms of this kind is in the postprocessing of the auditory error surface. Hollier et al. [17] uses an entropy measure of the error surface. MNB uses a hierarchical structure of integration over a range of time and frequency intervals. PESQ (Figure 2) furnishes the current state-of-the-art performance. PESQ performs integration in three steps, first over frequency, then over short-time utterance intervals, and finally over the whole speech signal. Different p values are used in the L_p norm integrations of the three steps. (PESQ also provides a delay compensation algorithm that is essential for quality measurement of

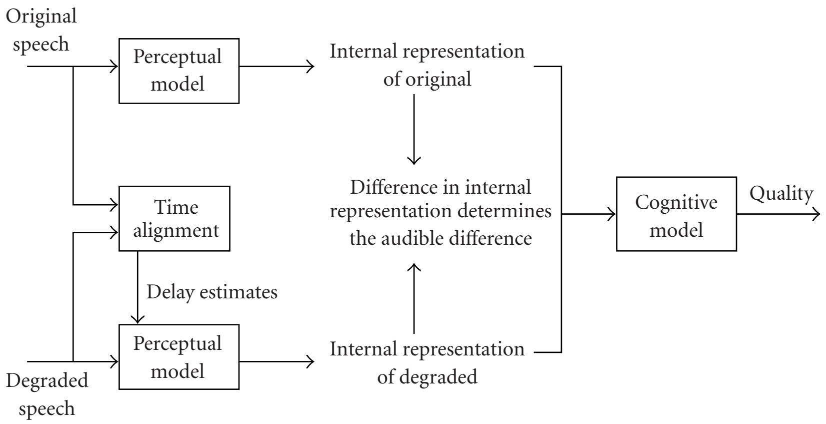

Original speech
Perceptual model
Internal representation of original
Time alignment
Difference in internal representation determines the audible difference
Cognitive model
Quality
Delay estimates
Perceptual model
Internal representation of degraded
Degraded speech

Figure 2: Schematic diagram of PESQ method [5].

voice packets that are subject to delay variation in the network.) The different methods of integration, though they may not resemble cognitive processes, achieve their respective degrees of effectiveness through using subjectively scored speech data to calibrate the mapping to estimated speech quality.

Subjects in MOS tests rate speech quality on the integer ACR scale of 1 to 5, with 5 representing excellent quality, and 1 representing the worst quality. The MOS is a continuous value based on averaging the listener's ACR scores. Ideally, the MOS obtained using a large and well-formed listener panel reflects the "true" mean opinion of the listener population. In practice, the measured MOS varies across tests, countries, and cultures. In subjective tests that use a different measure called DMOS [4], or degradation MOS, the subject listens to the original speech before scoring the degree of degradation of the degraded speech relative to the original. In MOS tests, a subject listens to a speech sample and chooses his/her opinion of its quality in a "categorical" sense, without first listening to a "reference" speech sample. The subject relies on his/her experience of speech quality to decide on the quality of the sample. Hence, single-ended algorithms are akin to MOS tests, while double-ended algorithms are akin to DMOS tests. Though most existing double-ended algorithms are designed to predict MOS, they may actually predict DMOS with better accuracy than MOS [18]. Relying on differences or distortions with respect to a "clean" signal alleviates the need to model "clean" speech in a normative sense. Nevertheless, distortions that are measurable on psychoacoustical scales do not necessarily contribute to perceived quality degradation. Speech signals can be modified in ways such that the modified signal can be distinguished from its original in a comparison test, but the modified signal would not be judged as degraded in a MOS test. Any "cognitive" processing ought to give no weight to differences that are measurable but do not affect the type of quality judgment that is predicted by the objective measurement. Existing double-ended algorithms do not have the intelligence to disregard such type of differences. The algorithms will predict a poorer quality for speech that has been transformed but not degraded. Consider the contrived example where an utterance is replaced with a different utterance of the same duration; the quality stays the same but the measured difference may be huge.

2.2. Statistical data mining

A major aim of this work is to use statistical data mining methods to find psychoacoustic features which most significantly correlate with quality judgment. Statistical data mining involves using statistical analysis tools to find underlying patterns or relationships in large data sets. Statistical data mining techniques have been applied to solve diverse problems in manufacturing quality control, market analysis, medical diagnosis, financial services, and so forth with much success. We consider two techniques in this paper: classification and regression trees (CART) [19] and multivariate adaptive regression splines (MARS) [20].

Suppose we have a response variable y and n predictor variables $x_1,\dots,x_n$. Suppose we observe N joint realizations of the response and predictor variables. Our observations can be modeled as

$$y = f(x_1,\dots,x_n) + \delta, \tag{1}$$

where δ represents a noise term. Our aim is to find a subset of predictor variables $\{x_{i_1},\dots,x_{i_m}\}$, $i_j \in \{1,\dots,n\}$, $j = 1,\dots,m$, $m \le n$, and a mapping $\hat{f}(x_{i_1},\dots,x_{i_m})$, such that $\hat{f}$ yields a good estimate of the response variable y.

2.2.1. CART

CART (classification and regression trees) [19] is a recursive partitioning algorithm. The domain D of the desired mapping is partitioned into M disjoint regions $\{R_m\}_1^M$. The partitioning process is recursive; new regions are generated by splitting regions that have been found so far. In conventional application of CART, the splitting is restricted to be perpendicular to the axis of the predictor variable chosen at each step of the recursion. This enables the splitting to be effected by answering a simple "yes" or "no" question on the predictor variable. The variable is chosen amongst $x_1,\dots,x_n$

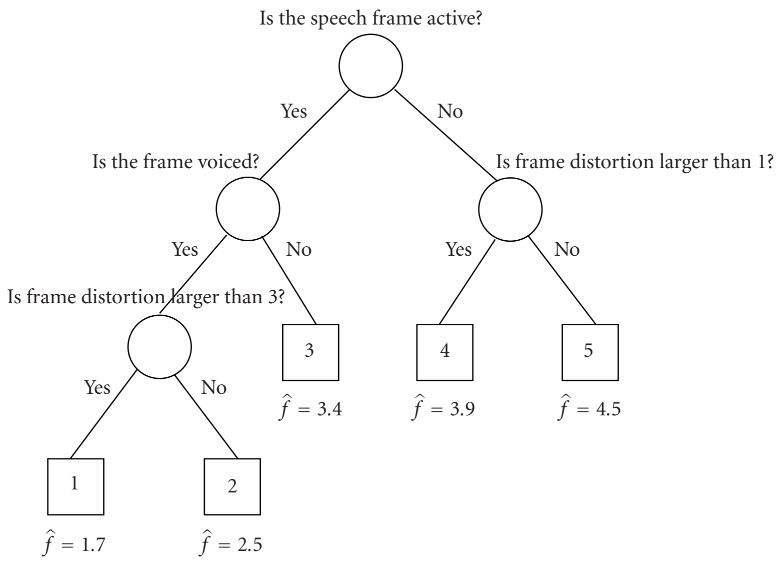


FIGURE 3: CART regression tree.

to minimize a splitting cost criterion. CART results are easy to interpret due to its simple binary tree representation. In Figure 3, a simplistic CART tree is shown, where circles represent internal nodes and rectangles represent leaf nodes. Each internal node in the tree is always split into two child nodes.

CART trees are designed in a two-stage process. First, an oversize tree is grown. The tree is then pruned based on performance validation, until the best-size tree is found. During tree growing, the next split is found by an exhaustive search through all possible single-variable splits at all the current leaf nodes. In CART regression, each region is approximated by a constant function

$$\hat{f}(\mathbf{x}) = a_m \quad \text{if } \mathbf{x} \in R_m. \tag{2}$$

The splitting cost criterion is the decrease in regression error resultant from the split. The regions generated by CART are disjoint, and the piecewise constant regression function $\hat{f}$ is discontinuous at region boundaries. This can lead to poor regression performance, unless the dataset is sufficiently large to support a large tree. Nevertheless, CART has been successfully used in classifying high-risk patients [19], quality control [21], and image vector quantization [22].

2.2.2. MARS

Multivariate adaptive regression spline (MARS) [20] was proposed as an improvement over recursive partitioning algorithms such as CART. Unlike CART, MARS produces a continuous regression function $\hat{f}$, and the regions of MARS may overlap. In MARS, $\hat{f}$ is constructed as a sum of M basis functions:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}), \tag{3}$$

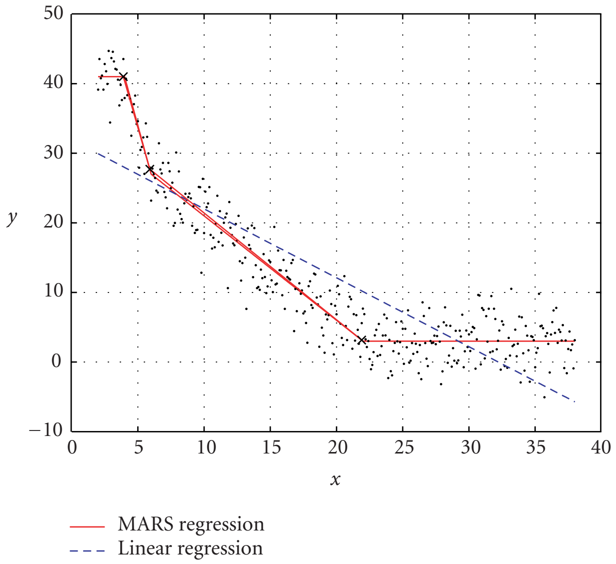


FIGURE 4: A MARS regression function with three knots (marked by ×).

where the basis function $B_m(\mathbf{x})$ takes the form of a truncated spline function. In Figure 4, a single-variable $\hat{f}$ with three "knots" is shown, where each knot marks the end of one region of data and the beginning of another. Compared to the linear regression function, the MARS regression function better fits the data.

Like CART, the MARS regression model is also built in two stages. First, an oversize model is built by progressively adding more basis functions. In the second stage, basis functions that contribute the least to modeling accuracy are progressively pruned. At each step in the model's growing phase, the best pair of basis functions to add is found

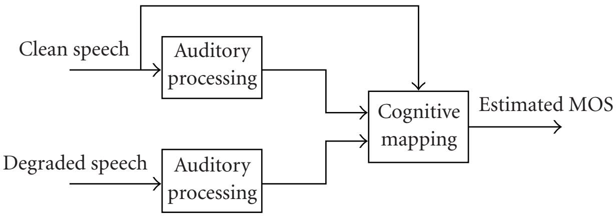


Figure 5: Algorithm architecture.

by an exhaustive search, similar to finding the best split during CART tree growing. MARS has been applied to predict customer spending and forecast recession [23], and predict mobile radio channels [24].

3. PROPOSED DESIGN METHOD

In the proposed method, double-ended measurement algorithms are designed based on the architecture depicted in Figure 5. Auditory processing (Figure 6) is first applied to both the clean speech and the degraded speech, to produce a subband decomposition for each signal. The subband decomposed signals and the clean speech signal are input to the cognitive mapping module (Figure 7), where a distortion surface is produced by taking the difference of the two subband decompositions. A large pool of candidate feature variables is extracted from the distortion surface. MARS and/or CART is applied to sift out a small set of predictor variables from the pool of candidate variables, while progressively constructing and optimizing the regression mapping $\hat{f}$. This mapping replaces the statistical mining block in Figure 7 upon completion of the design.

The auditory processing modules decompose the input speech signals into power distributions over time frequency and then convert them to auditory excitations on a loudness scale. The cognitive mapping module interprets the differences (distortions) between the auditory excitations of the clean and the degraded speech signals. In effect, the cognitive module "integrates" the distortions over time and frequency to arrive at a predicted quality score. We make the simple observation that "distortions are not created equal." An isolated large distortion event is likely to be cognitively distinct from small distortions that are widely diffused over time frequency, though the small distortions may integrate to a substantial amount. The latter kind of distortion may be less annoying than the former kind. We take an agnostic view of how human cognition weighs the contributions from different types of distortions. The approach we take is to create a plethora of "contexts" under which distortion events occur. Distortions with the same context are integrated to a value which we call a "feature." Straightforward root-mean-square (RMS) integration is used to compute the feature value. Each context gives rise to one *candidate* feature, so that there are as many candidate features as the number of contexts. From the pool of candidate features, data mining techniques are used to find a small subset of features and the best way to combine them to estimate the speech quality. The modules are described next in Sections 3.1 and 3.2. Detailed design considerations and justifications of the modules then follow in Section 3.3, and finally computational complexity is considered in Section 3.4.

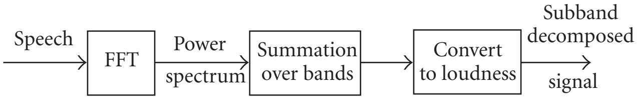


Figure 6: The processing steps in an auditory processing module.

3.1. *Auditory processing*

A block diagram of auditory processing is depicted in Figure 6. Human auditory processing of acoustic signals is commonly modeled by signal decomposition through a bank of filters whose bandwidths increase with filter center frequency according to the bark or critical-band scale [8]. A typical realization of this model employs roughly 17 filters or spectral bands to cover the telephone voice channel. In our experiments, we found that 7 bands, each with bandwidth of about 2.4 bark, strike a good balance between prediction performance and sensitivity to irrelevant variations in the input data (for further elaboration, see Section 3.3.1). In our scheme, the speech signal is partitioned into 10-millisecond frames. For each frame, a 128-point power spectrum is calculated by applying FFT to a 128-point Hanning-windowed signal segment centering on each frame. The spectral power coefficients are grouped into 7 bands. The coefficients in each band are summed, to produce altogether 7 subband power samples. The samples are converted to subjective loudness scale using Zwicker's power law [8]:

$$L(f) = L_0\left(\frac{E_{TQ}(f)}{s(f)E_0}\right)^k\left[\left(1 - s(f) + \frac{s(f)E(f)}{E_{TQ}(f)}\right)^k - 1\right], \quad (4)$$

where the exponent $k = 0.23$, $L_0 = 0.068$, E_0 is the reference excitation power level, $E_{TQ}(f)$ is the excitation threshold at frequency f, $E(f)$ is the input excitation at frequency f, and $s(f)$ is the threshold ratio.

3.2. *Cognitive mapping*

The "cognitive mapping" module comprises functional blocks as depicted in Figure 7. The decomposed clean and degraded speech signals from the auditory processing modules are first subtracted to obtain their absolute difference, which is called the "distortion". The distortion over the whole speech signal can be organized into a two-dimensional array, representing a distortion surface over time frequency. A goal of the cognitive mapping is to aggregate cognitively similar distortions through segmentation and classification (elaborated below). The perceptually significant aggregated distortions are found using data mining. The statistical data mining block in Figure 7 is present during the design phase of the cognitive mapping block. Once the design is completed,

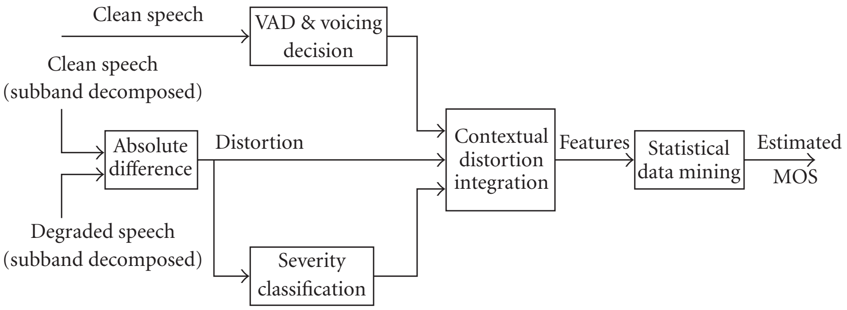


Figure 7: Cognitive mapping.

the block is replaced by a simple mapping block. The mapping (the aforementioned $\hat{f}$) is computationally simple, as can be seen from the example presented in Appendix B.

3.2.1. *Time segmentation*

The clean speech signal is processed through a voice activity detector (VAD) and then a voicing detector. Each 10-millisecond speech frame is thereby labeled as either "background," "active-voiced," or "active-unvoiced." We use the VAD algorithm from ITU-T G.729B [25], omitting the comfort noise generation part of the algorithm. More recent VAD algorithms such as that in the AMR codec [26] may also be used to advantage. The purpose of the segmentation is to separate the different types of speech frames so that they can exert separate influence on the speech quality estimate. The advantage of such segmentation is suggested in [27], where performance was improved using clustering-based segmentation.

3.2.2. *Severity classification*

The total distortion of each frame is classified into different severity levels. The aim is to sift out the significant distortion events. Different forms of classifiers can be used. We have experimented with simple thresholding, CART classification [19], and Gaussian mixture density modeling. Based on our simulation results, we have found that a simple classification scheme, thresholding the average frame distortion, suffices to produce most of the benefit. Results presented below are based on thresholding to 3 severity levels, which we call low, medium, and high distortion severity. In [28], fixed thresholding of frame energy is shown to provide performance gain. Gains obtained from classification and segmentation are discussed in Section 3.3.2.

3.2.3. *Context and aggregation*

The speech signal now has a time-frequency representation, with a distortion sample in each time-frequency bin. Each sample is labeled according to its frequency band index, time-segmentation type, and severity level. Contexts are created by combining label values. For instance, the above segmentation and classification creates $7 \times 3 \times 3 = 63$ distinct values. The distortion samples that have the same composite-label value belong to the same context, which is named after the composite-label value. By associating a context with each distinct composite-label value, we form 63 distinct contexts. Each context contributes one feature variable to the candidate feature pool to be mined. The value of a feature is obtained via root-mean-square integration of the distortion samples in the context, normalized by the number of frames in the speech signal. Thus, each context establishes a specific class of distortion, and contributes to data mining a feature variable which captures the level of the distortion in that class. The feature variables are defined in Appendix A. As an example, the variable U_B_2_0 captures the integrated distortion of the context: unvoiced frame, subband 2, and low severity (level 0). We assume that the lengths of the speech signals are no more than several seconds so that recency effects can be ignored. Recency effects can be accounted for by introducing forgetting factors.

3.2.4. *Feature pool*

Additional contexts are defined in order to create a "rich" pool of candidate features for mining. Besides labeling each frequency subband with its natural subband index, each subband is also labeled with the rank order obtained by ranking the 7 distortions in a frame in order of decreasing magnitude. Thus, a candidate feature has either a natural or ordered subband index. Rank ordering the subband distortions as well as classifying frame-level distortions based on severity create contexts that capture distortions independent of specific time-frequency locations, but dependent on the absolute or relative level of distortion severity. This is hypothetically justifiable by the nature of the quality judgment process, and helps the data mining algorithm to pick out cognitively significant events.

Additional contexts are also created by omitting some labels such as the severity level. These contexts are the 7 subbands, in natural or ordered index, for each of the 3 time-segmented frame classes, without severity classification; altogether there are $7 \times 3 = 21$ such contexts (whose feature variables are listed in Appendix A as T_B_*b* and T_O_*b*). We also include weighted mean and root-mean distortions, probability of each frame type, and the lowest-frequency-band

and the highest-frequency-band energy of the clean speech frames, to produce a pool totaling 209 candidate features, as listed in Appendix A. The weighted mean of the 7 subband distortions is calculated using the weights [29]

$$w_i = \begin{cases} 1.0 & \text{for } 0 \le i \le 4, \\ 0.8 & \text{for } i = 5, \\ 0.4 & \text{for } i = 6. \end{cases} \tag{5}$$

The pool of candidate features is redundant for the purpose of quality estimation. A brute force approach to finding the best subset of features to use would entail examining $2^{209} - 1$ possible subsets, a clearly impossible task. Yet the success of our approach crucially depends on finding a small subset of features that are good for quality estimation. We resort to data mining techniques to perform this task. The effectiveness of the techniques and performance of their designs are assessed experimentally in Section 4.

3.3. Feature design and selection

In this section, we present some design justifications.

3.3.1. Number of subbands

We first experimented with using 22 subbands, with each band roughly three-quarter-bark wide. Using CART for regression, we found that roughly one out of every three bands was selected. Therefore, we conjectured that we could group the distortions over 22 subbands into a smaller set of 7 subband distortions, to achieve a better tradeoff between retaining relevant spectral information and easy generalization. In a similar rein, reduced spectral resolution was found to improve the accuracy of speaker-independent speech recognition [30]. The 2.4-bark bandwidth in our frequency decomposition can also be compared with the 3–3.5 bark critical distance between vowel formant peaks [31].

3.3.2. Design of segmentation and severity classification

In this section, we show the improvements on speech quality estimation due to using segmentation and severity classification. Estimation performance is assessed using the correlation R and root-mean-square error (RMSE) ϵ between the subjective MOS x_i and objective MOS y_i. Pearson's formula gives

$$R = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^N (x_i - \bar{x})^2 \sum_i^N (y_i - \bar{y})^2}}, \tag{6}$$

where $\bar{x}$ is the average of x_i, and $\bar{y}$ is the average of y_i. RMSE is calculated using

$$\epsilon = \sqrt{\frac{\sum_{i=1}^N (x_i - y_i)^2}{N}}. \tag{7}$$

The performance results exhibited in Table 1 are based on designing a MARS model for a speech database. As we can see, time segmentation alone provides some improvement. An interesting phenomenon is that distortion severity classification alone does not result in any improvement. However, a large improvement is obtained by combining segmentation and classification. We attribute this phenomenon to the different significance of a given distortion level across the three types of speech frames: inactive (background noise), voiced, and unvoiced. The signal contents of the three frame types are perceptually very distinct. We expect each type of contents to condition the perception of distortion in a certain characteristic fashion. Separating the distortions according to the frame types allows the distortions to be weighed differently for each type.

Table 1: Performance with different combinations of segmentation and severity classification.

	Correlation R	RMSE ϵ
No segmentation or classification	0.906	0.306
Segmentation only	0.927	0.273
Severity classification only	0.896	0.339
Segmentation and severity classification	0.977	0.155

For feature definition, we also compared between using (i) the number of distortion samples in a severity class, normalized by the number of frames in a speech file, versus (ii) RMS integration of the distortion samples in a severity class. The latter was found to provide better performance.

3.3.3. Feature selection

In this section, we acquire a sense of the features selected by MARS by perturbing a MARS designed model. The "Original" column in Table 2 lists the variables of the model being perturbed, in order of decreasing importance. Variable importance is determined by the amount of reduction in prediction error provided by the variable, relative to the greatest reduction amount achieved amongst all variables. Hence, the variable that results in the largest prediction error reduction has importance 100%, and its amount of error reduction is used as reference. The importance of other variables is calculated as the percentage of their prediction error reduction relative to the reference.

An inspection of the "Original" list naturally raises the question of why some of the variables are important. For example, I_P_VUV, the ratio of the number of inactive frames to the number of active frames, is rated most important. Moreover, low-rank subband distortion variables U_O_4, U_O_3, U_O_5, and U_O_6_1 are included in the model, and yet the high-rank subband distortions of the same unvoiced frame type are not included. To address these questions, we removed the above feature variables from the candidate pool and redesigned the model. The resultant features are listed in the "Modified" column of Table 2. We see from this list that I_P, the fraction of inactive frames, is rated more important than before. I_P and I_P_VUV provide different encoding of the same information, but I_P and I_P_VUV are not linearly related. Also, in lieu of the omitted low-rank subband variables for the unvoiced frames, the high-rank subband variable U_O_0 is brought into the modified model. Thus, we see

Table 2: Variable importance list for feature selection investigation. Original: list generated using the full feature pool. Modified: list generated after trimming the feature pool.

Rank	Original		Modified		Rank	Original		Modified	
	Variable	Import.	Variable	Import.		Variable	Import.	Variable	Import.
1	I_P_VUV	100.000	V_B_2	100.000	11	V_P_VUV	34.706	I_B_1_0	29.661
2	V_B_5	68.859	I_P	68.556	12	I_WM_1	33.749	I_WM_1	24.958
3	V_B_2	68.051	V_B_2_2	58.165	13	I_B_1_0	33.033	V_B_0_1	22.584
4	V_B_2_2	47.966	V_O_0	49.106	14	V_B_3	32.877	V_WM_0	15.747
5	U_P_VUV	47.214	REF_1	41.957	15	U_B_2	24.440	V_B_3	15.665
6	V_O_0	42.583	I_B_0	39.036	16	V_B_0_1	23.568	V_RM_0	15.339
7	I_B_0	42.382	V_P	37.517	17	U_O_3	21.882	V_B_5_1	11.970
8	REF_1	41.220	I_B_2	36.124	18	V_O_4	18.121	U_O_0	10.877
9	I_P	41.014	V_B_5	35.681	19	U_O_5	15.959	I_O_0	9.818
10	U_O_4	36.489	V_P_VUV	35.255	20	U_O_6_1	14.487	—	—

that both the information captured in a variable as well as the manner of encoding of the information in the variable affect its importance. A rich candidate pool should convey a variety of information as well as information encoding. MARS consistently picks out from the available feature variables, the ones with the most relevant information and the best encoding. The original model, drawn from a richer pool, is preferred over the modified model. The original model provides root-mean-square prediction error (RMSE) of 0.3902 and 0.3844 on the 90% training database and 10% test database, respectively. (Databases and performance assessment are discussed in the next section.) For the same databases, the modified model achieves RMSE of 0.3968 and 0.4318, respectively.

3.4. Complexity

The computational complexity of the algorithms designed using the proposed approach is mainly attributable to the auditory processing modules and to feature extraction processing in the cognitive module. While the design of the mapping from features to the MOS estimate is somewhat involved, the actual processing needed to realize the mapping once it is designed is simple. As the purpose of this paper is to study the application of data mining techniques to design speech quality measurement algorithms, we offer below a rough guide of the algorithm complexity. The actual complexity in specific applications will vary with the details of the features selected. Moreover, as with other measurement algorithms (see, for example, [32]), algorithm complexity may be reducible without seriously degrading the estimation accuracy. Such pursuit of complexity reduction is left to future study.

The complexity of auditory processing in the designed algorithms is no greater than that of the auditory processing component in PESQ. A somewhat lower complexity is obtained in our case by using fewer subbands. RMS integration of distortion samples to compute the values of the features employed in the data-mining designed mapping has a roughly similar complexity to the L_p integrations performed in PESQ. Our use of squared integration throughout, as opposed to using several different values of p in PESQ, lowers the integration complexity. Computation of the mapping function (see Appendix B for an example), done only once for the whole speech file, has relatively negligible complexity. Severity classification also has negligible complexity. The segmentation functionalities, VAD and voicing decision, are commonly found in speech coders and other speech processing applications. We have used the VAD algorithm in ITU-T G.729B [25], omitting its comfort noise generation functionality. We estimate that the segmentation functionalities require no more than 20% of the processing time of the ITU-T G.729 speech codec. The processing time of PESQ is roughly 2.8 times that of G.729. We note that PESQ provides additional functionalities such as variable delay compensation. Hence, a speech quality estimator using an algorithm designed using the proposed approach while providing a similar suite of functionalities as PESQ would incur a 7% higher complexity than PESQ. As this is a conservative upper bound, we believe complexity implementations lower than PESQ are readily achievable.

4. EXPERIMENT RESULTS

The effectiveness of the data mining approach is demonstrated experimentally with actual designs. We compare the performance of the algorithms designed using our method to the current state-of-the-art algorithm in voice quality estimation, PESQ. Below, we first introduce the speech databases used for the experiments. Then we compare the designs obtained using different data mining techniques, namely CART, hybrid CART-MARS, and MARS. We finally focus on the method that offers the best performance: MARS design using cross validation. The greatest difference between our designed algorithms and PESQ is in the cognitive mapping part; thus, the comparisons below can be regarded as evaluating different cognitive mappings.

4.1. Speech databases

The speech databases used in our experiments are listed in Table 3. They include the 7 multilingual databases in ITU-T P-series Supplement 23 [33], two wireless databases (IS-96A and IS-127 EVRC), and a mixed wireline-wireless database [18]. We combine the 10 databases into a global database for algorithm design. There are altogether 1760 degraded-speech files in the global database.

Table 3: Properties of the speech databases used for experiments.

Database	Language	No. of files	Minimum MOS	Maximum MOS	Average MOS	MOS spread	MOS std. error
ITU-T Supp23 Exp1A	French	176	1.000	4.583	3.106	0.781	0.148
ITU-T Supp23 Exp1D	Japanese	176	1.000	4.208	3.666	0.701	0.158
ITU-T Supp23 Exp1O	English	176	1.208	4.542	3.050	0.822	0.155
ITU-T Supp23 Exp3A	French	200	1.292	4.833	3.226	0.732	0.152
ITU-T Supp23 Exp3C	Italian	200	1.083	4.833	2.950	0.896	0.152
ITU-T Supp23 Exp3D	Japanese	200	1.042	4.417	2.331	0.737	0.155
ITU-T Supp23 Exp3O	English	200	1.167	4.542	2.782	0.772	0.187
Wireless IS-127 EVRC	English	96	2.250	4.500	3.427	0.500	0.340
Wireless IS-96A	English	96	1.625	3.875	2.760	0.451	0.341
Mixed	English	240	1.090	4.610	3.200	0.728	n.a.

The three Exp1x databases in ITU-T Supp23 contain speech coded using the G.729 codec, singly or in tandem with one or two other wireline or wireless standard codecs, under the clean channel condition. Also included are single-encoded speech using these standard codecs. The four Exp3x databases contain single- and multiple-encoded G.729 speech under various channel error conditions (BER 0%–10%; burst and random frame erasure 0%–5%) and input noise conditions (clean, street, vehicle, and hoth noises at 20 dB SNR).

The wireless IS-96A and IS-127 EVRC (Enhanced Variable Rate Codec) databases contain speech coded using the IS-96A and IS-127 codecs, respectively, under various clean and degraded channel conditions (forward FER 3%, reverse FER 3%), with or without the G.728 codec in tandem, and MNRU (modulated noise reference unit) conditions of 5–25 dB. The mixed database [18] contains speech coded with a variety of wireline and wireless codecs, under a wide range of degradation conditions: tandeming, channel errors (BER 1%–3%), and clipping (see [18] for more details). All databases include reference conditions such as speech degraded by various levels of MNRU.

The range of the MOSs in each database is determined by its mix of test conditions. The range is characterized in Table 3 by the maximum, minimum, average, and "spread", which is the standard deviation of the MOSs around the average. The imprecision of the subjective MOS is characterized by its standard error ("MOS std. error" in Table 3, which is determined by the number of listeners who participated in the subjective test). The RMSE of the objective scores can be assessed no better than the standard error of the subjective scores used to benchmark the accuracy. Moreover, the measurement accuracy of algorithms trained using a database is also limited by the imprecision of its subjective scores. Note that "No. of files" in Table 3 refers to the number of speech files that are subjectively scored; the "clean original" speech files are not counted.

The designs presented in this paper are based on the above databases which cover a range of waveform codecs, wireline and wireless LPAS [14] codecs, and a range of codec tandeming and channel error conditions, and input background noise conditions. Additional impairments that can be found in telephone connections but are not currently covered by our databases include echo, variable delay, tones, distortions due to harmonic or sinusoidal coders and due to music and artificial speech, and so forth the reader can also consult [5] for its list of transmission impairments. The proposed design method is highly automated and should scale well with the amount of database material available for design (see Section 4.6).

4.2. CART results

We first experimented with using CART for mining, motivated by the fact that CART results are easier to interpret than MARS results, and CART can be regarded as a special case of MARS. For CART mining, we randomly assigned 90% of the global database to a training data set and the rest to a test data set. The tree-growing phase uses the training set, and the tree-pruning phase uses the test set to select the best-size tree, that is, the one that gives the lowest regression error on the test data. The CART-designed tree has 38 leaf nodes. The performance scores are $R = 0.8861$ and $\epsilon = 0.3734$ on the training set, and $R = 0.7627$ and $\epsilon = 0.5098$ on the test set. The large difference in RMSE values between training and testing indicates that the designed CART tree does not generalize well. For PESQ, we use the PESQ-LQ mapping suggested in [34] to obtain $R = 0.8170$ and $\epsilon = 0.4705$ on the global undivided set, $R = 0.8198$ and $\epsilon = 0.4700$ on the training set, and $R = 0.7939$ and $\epsilon = 0.4744$ on the test set. It appears that CART regression trees cannot outperform PESQ.

4.3. Hybrid CART-MARS results

By inspecting the variables mined using CART, we expect them to be perceptually important. The poor performance might be due more to the aforementioned limitations of CART in regression, rather than to the feature selection. Thus, we experimented with using MARS to circumvent the limitations of CART. Below, we present the results from two hybrid CART-MARS schemes.

The first hybrid CART-MARS method uses CART to prescreen features from the feature candidate pool. The feature variables selected by CART are used as a smaller feature candidate pool for MARS model building. In this method, CART is used only during model design; the final model is constructed solely by MARS. The performance obtained, $R = 0.8501$ and $\epsilon = 0.4242$ on the training set, and $R = 0.8233$ and $\epsilon = 0.4379$ on the test set, is better than PESQ and CART regression.

Table 4: MARS model selection as a function of DS using 10-fold cross-validation.

DS	N	M	Training			Testing		
			R	ϵ	%	R	ϵ	%
3	78	125	0.9261	0.3025	35.6	0.8403	0.4409	7.1
6	47	66	0.9055	0.3402	27.6	0.8527	0.4200	11.5
10	21	39	0.8880	0.3685	21.6	0.8550	0.4164	12.2
15	20	25	0.8756	0.3872	17.6	0.8530	0.4182	11.8
20	19	21	0.8707	0.3941	16.1	0.8546	0.4156	12.4
25	16	18	0.8652	0.4019	14.5	0.8502	0.4223	11.0

The second hybrid CART-MARS method is similar to the method used in [35]. In [35], the feature candidate pool for MARS mining is augmented by the "leaf-node index" obtained from a CART tree. We improve on the method by adding the CART regression output variable, instead of the node index variable, to the candidate feature pool. The augmented candidate pool is used for MARS model building. In this method, if the CART output were incorporated into the MARS model, feature extraction for the model would also include computation as prescribed by the CART tree. Indeed, an inspection of the variable importance list found that the CART tree output is the most important feature variable selected. The performance obtained, $R = 0.9108$ and $\epsilon = 0.3326$ on the training set, and $R = 0.8231$ and $\epsilon = 0.4423$ on the test set, is also better than PESQ and CART regression. The larger difference between training and testing RMSE in this "augmentation" method, in comparison with the earlier "prescreening" method, suggests that the "prescreening" method is more robust.

Although both hybrid CART-MARS methods outperform PESQ and CART, they are inferior to the MARS model of Section 3.3.3 on the test set. In the rest of this paper, we present detailed results based on using MARS alone, as MARS tends to offer the best performance. For the application in [35], a hybrid CART-MARS scheme provides better performance than CART or MARS alone. Thus, we should not eliminate the possibility of some hybrid schemes outperforming MARS-only schemes.

4.4. MARS model selection via cross-validation

Picking the size of the regression model is a crucial step in the design. The size of the model designed using MARS is a function of M, the number of basis functions in (3). For linear spline basis functions, two real parameters are associated with each function, the "knot" and the linear combination weight. (Please refer to the example in Appendix B.) Thus, the number of optimized parameters, $2M$, is a useful measure of model size. A large model yields low regression error, but the model is highly biased towards the training data and exhibits large variance over unseen data. On the other hand, a small model might omit some important features necessary for high measurement accuracy. In Friedman's original MARS design [20], a penalty term controlled by a "degree of smoothness" (DS) parameter is used in the criterion function to penalize the increased variance due to large model size. Larger DS results in more basis functions taken out during the pruning phase. Friedman's design method does not incorporate validation of the model through testing with data not used in model building. We improve on Friedman's design by using cross-validation to select the model size.

In conventional model design, available data is split into a training set and a test set. The model is built on the former, and validated on the latter. However, when the amount of available data is small, as in our case, we ought to use all the data for model building. Using a small sample to design and validate can be achieved by n-fold cross-validation [36]. The results presented below are based on $n = 10$-fold cross-validation. The global database is randomly divided into 10 data sets with almost equal size. Training and testing is performed 10 times. Each time, one of the data sets serves as the test set, and the remaining 9 data sets combined serve as the training set. Each data set serves as a test set only once. For each training-test set combination, a series of MARS models corresponding to various DS values are constructed using the training set. The 10 R and ϵ values obtained for each DS value are averaged to obtain the cross-validation R and ϵ values; separate averages are obtained from the training and test sets. Finally, the DS value corresponding to the best cross-validation performance is used to build the desired MARS model using the entire global database.

Table 4 shows the cross-validation performance results for a series of MARS models obtained using different values of DS. Both training and test results are shown, with N denoting the average number of distinct feature variables used in the cross-validation models, M the average number of basis functions, and % the average percentage reduction in ϵ compared to PESQ. From Table 4, we pick the best DS value for designing our final model. We see that for DS = 20, the RMSE reduction is the largest, and the discrepancy between the training and test performance is the smallest. Thus, the final model is built using the global database, with DS = 20.

The resultant "global model" has $N = 21$ feature variables and $M = 24$ basis functions. The variables and their importance are listed in Table 5, and the MARS regression function and its basis functions are given in Appendix B. We see from Table 5 that the most important variables and most of the variables are related to voiced frames. The overall trend is that features from voiced frames are treated as more important than those from unvoiced and inactive frames. This is consistent with the fact that the great majority of active speech frames are voiced and that human perception is more sensitive to distortion of the spectral envelopes of voiced frames than unvoiced frames. The most important variable V_RM, the root-mean distortion of voiced frames, is akin to

Table 5: Variable importance ranking for the global model.

Rank	Variable	Importance	Rank	Variable	Importance	Rank	Variable	Importance
1	V_RM	100.00	8	I_B_0	39.782	15	U_O_4	24.530
2	I_P	76.280	9	I_B_1_0	37.531	16	U_O_5	22.092
3	V_B_2_2	56.375	10	V_O_0	37.143	17	V_P_1	21.172
4	REF_1	50.151	11	V_B_2	36.379	18	U_B_4	19.639
5	V_P	44.425	12	V_O_5	33.179	19	V_B_0_1	17.459
6	V_O_0_2	41.897	13	I_WM_1	31.807	20	U_O_6_1	17.048
7	V_P_VUV	40.472	14	V_P_2	25.562	21	I_B_5	15.592

Table 6: MARS model performance on the 10 speech databases: variation over samples.

Database	Language	Correlation R		RMSE ϵ		Percentage
		Proposed Method	PESQ	Proposed Method	PESQ	reduction in ϵ (%)
ITU-T Supp23 Exp1A	French	0.8753	0.8498	0.3909	0.4507	13.3
ITU-T Supp23 Exp1D	Japanese	0.9141	0.8725	0.3988	0.5893	32.3
ITU-T Supp23 Exp1O	English	0.8998	0.9164	0.3581	0.3616	1.0
ITU-T Supp23 Exp3A	French	0.8480	0.8199	0.4327	0.5482	21.1
ITU-T Supp23 Exp3C	Italian	0.9099	0.8935	0.4048	0.4499	10.0
ITU-T Supp23 Exp3D	Japanese	0.8728	0.8965	0.4127	0.5366	23.1
ITU-T Supp23 Exp3O	English	0.8757	0.8857	0.3749	0.4222	11.2
Wireless EVRC	English	0.6364	0.5522	0.3952	0.4359	9.3
Wireless IS-96A	English	0.5786	0.4562	0.3845	0.4282	10.2
Mixed	English	0.8771	0.8732	0.3496	0.4083	14.4
Average	—	—	—	—	—	14.6

the logarithmic spectral distortion that speech spectral quantizers are generally designed to minimize [14]. V_B_2_2 is the RMS distortion in subband 2 of the voiced frames that have the highest severity of frame distortion. Subband 2 covers the frequency region where the long-term power spectrum of speech peaks. V_O_0_2 is the RMS distortion in the highest-distortion subband of the voiced frames that have the highest severity of frame distortion; in effect, V_O_0_2 measures the intensity of peak distortions. The selection of V_B_2_2 and V_O_0_2 suggests that speech quality perception is strongly dependent on prominent spectral regions and distortion events. The variables I_P, V_P, and V_P_VUV, which measure the relative amount of specific frame types, and REF_1, which measures the level of high-frequency loudness in the reference signal, serve to adjust the regression mapping. For instance, in Appendix B, we see that the predicted quality value is raised when the fraction of inactive frames is above 0.27, and is decreased when the fraction drops below 0.27.

4.5. Database results

We apply the global model to the individual databases listed in Section 4.1. We report performance results in two formats: variation over samples (VOS) in Table 6, and variation over conditions (VOC) in Table 7. In VOS, the correlation and RMSE between the objective and subjective MOS of each sample is reported. A "sample" refers to a pair of speech files used for quality calculation: the speech file that was played to the listener panel, and the "clean" original version of the speech that was played. For VOC, the subjective MOSs for the speech files within the same test condition are first averaged together.

The objective MOSs are also likewise grouped and averaged. Then, R and ϵ are calculated between the per-condition averaged subjective and objective MOSs, over all conditions in the database. The VOS results better reflect performance in voice quality monitoring applications [3]. The VOC results are more appropriate for codec or transmission equipment evaluation. To the best of our knowledge, all the performance results that have been reported in the literature for PESQ by its inventors use the VOC format. The results for PESQ are based on using the PESQ-LQ 3rd-order regression polynomial specified in [34]. The results in Tables 6 and 7 show that the global model provides an average reduction in RMSE ϵ of 14.6% and 21.4%, for VOS and VOC averaging, respectively.

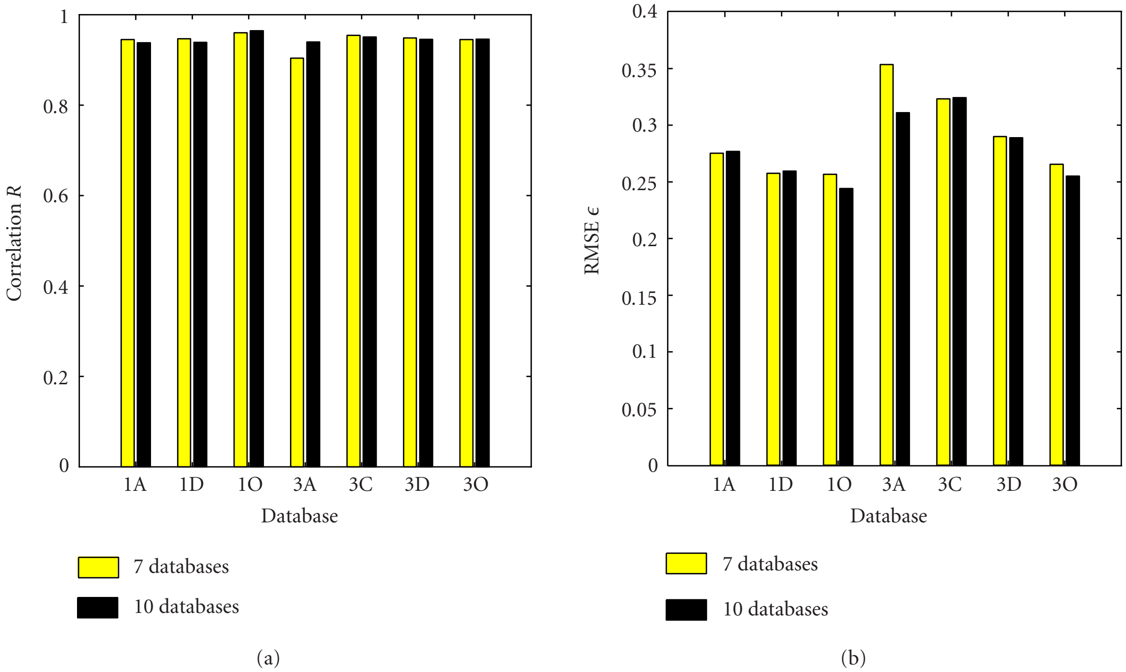
We adopt the simple model proposed in [7] to help us interpret the relationship between the R and ϵ values in Tables 6 and 7; the model is modified with the addition of a bias term. Accordingly, R and ϵ satisfy the following relationship:

$$\epsilon^2 = \sigma^2(1 - R^2) + \sigma^2_{\text{MOS}} + b^2, \tag{8}$$

where σ^2 and σ^2_{MOS} are the "MOS spread" and "MOS std. Error" in Table 3, respectively, and b is systematic bias. The equation states that ϵ^2 is the sum of unexplained variance in the estimation model, MOS estimation error due to limited number of listeners, and bias error between subjective and objective MOSs. In comparing estimation algorithms using the same databases, σ^2_{MOS} is an irreducible noise term affecting all the algorithms equally. Tables 6 and 7 show that PESQ produces large ϵ values on databases Exp1D, Exp3A,

TABLE 7: MARS model performance on the 10 speech databases: variation over conditions.

Database	Language	Correlation R		RMSE ϵ		Percentage reduction in ϵ (%)
		Proposed Method	PESQ	Proposed Method	PESQ	
ITU-T Supp23 Exp1A	French	0.9381	0.9343	0.2769	0.3609	23.3
ITU-T Supp23 Exp1D	Japanese	0.9391	0.9539	0.2595	0.5136	49.5
ITU-T Supp23 Exp1O	English	0.9644	0.9566	0.2441	0.2705	9.8
ITU-T Supp23 Exp3A	French	0.9400	0.8776	0.3109	0.4743	34.5
ITU-T Supp23 Exp3C	Italian	0.9508	0.9455	0.3243	0.3441	5.8
ITU-T Supp23 Exp3D	Japanese	0.9455	0.9452	0.2888	0.4785	39.6
ITU-T Supp23 Exp3O	English	0.9459	0.9254	0.2551	0.3522	27.6
Wireless EVRC	English	0.8224	0.8116	0.2139	0.2176	1.3
Wireless IS-96A	English	0.6323	0.6203	0.2371	0.2250	−5.4
Mixed	English	0.9364	0.9188	0.2438	0.3366	27.6
Average	—	—	—	—	—	21.4



(a)

(b)

FIGURE 8: Comparison of MARS model performance between training on the 7 ITU-T databases and on all 10 speech databases. (a) Correlation and (b) RMSE results are shown for variation over conditions.

and Exp3D, even though R is quite high for databases Exp1D and Exp3D. According to (8), the large ϵ values can be due to bias errors, which we attribute to biases between individual databases and the global database. The MARS model is able to adjust for individual databases, thus reducing the bias component.

4.6. *Scalability*

It is highly desirable to be able to design models that can scale with the amount of data available for learning. Also, new forms of speech degradations arise as a result of new transmission environments, new speech codecs, and so forth. The data mining approach enables designing best-size models for a given amount of learning data, and adapting to new learning data. To demonstrate the scalability of the proposed method, we created a smaller global database comprising only the seven ITU-T databases. New MARS models with different DS values were designed using the new global database. In Figure 8, we compare the performance of the new model, with DS $=$ 20, $N = 13$, and $M = 15$, to that of the larger global model designed earlier. The results are for VOC; the results for VOS are similar. One might expect the MARS model designed for the global database to be "diluted" and hence less effective than the new model designed for the seven ITU-T databases. However, we see that the two models provide about the same level of performance. In fact, it is somewhat surprising that the global model furnishes 14% lower RMSE than the more tuned seven-database MARS model. Thus, the proposed method appears to scale well with the amount of learning data, and suggests favorably the possibility of large-scale (semi-)automated, online model (re-)training.

5. CONCLUSION

We have proposed an approach to design objective speech quality measurement algorithms using statistical data mining methods. We have examined various methods of using CART and MARS to design novel objective speech quality measurement algorithms. The methods select feature variables from a large pool to form speech quality estimation models. We have obtained designs that outperform the state-of-the-art standard PESQ algorithm in our databases. The variables forming the models are found to be perceptually significant, and the methods offer some insights into the relative importance of the variables. The designed algorithms are computationally simple, making them suitable for real-time implementation. The best performing algorithm was designed using MARS.

We also showed that the proposed design method can scale with the amount of learning data. The experience learned from building training-based systems such as speech recognizers suggests using that the performance of the algorithms designed using our approach can be substantially improved with large-scale training, offline or online. The algorithms also show promise for further optimization and complexity reduction. The design approach can be extended to other media modalities such as video.

APPENDICES

A. FEATURE VARIABLE DEFINITIONS

The feature variables are defined below. The first letter, denoted by T in a variable name, gives the frame type: T = I for Inactive, T = V for Voiced, and T = U for Unvoiced. The subband index is denoted by b, with $b \in \{0, \dots, 6\}$ indexing from the lowest to the highest frequency band if the index is natural, or from the highest to the lowest distortion if the index is rank-ordered. The frame distortion severity class is denoted by d, with $d \in \{0, 1, 2\}$ indexing from lowest to highest severity. With the above notations, the feature variables are as follows.

(i) T_P_*d*: fraction of T frames in severity class *d* frames.
(ii) T_P: fraction of T frames in the speech file.
(iii) T_P_VUV: ratio of the number of T frames to the total number of active (V and U) speech frames.
(iv) T_B_*b*: distortion for subband *b* of T frames, without distortion severity classification, for example, I_B_1 represents subband 1 distortion for inactive frames.
(v) T_B_*b*_*d*: distortion for severity class *d* of subband *b* of T frames, for example, V_B_3_2 represents distortion for subband 3, severity class 2, of voiced frames.
(vi) T_O_*b*: distortion for ordered subband *b* of T frames, without severity classification, for example, U_O_3 represents ordered-subband 3 distortion for unvoiced frames, without distortion severity classification.
(vii) T_O_*b*_*d*: distortion for distortion class *d* of ordered subband *b* of T frames, for example, U_O_6_1 represents distortion for severity class 1 of ordered-subband 6 of unvoiced frames.
(viii) T_WM_*d*: weighted mean distortion for severity class *d* of T frames.
(ix) T_WM: weighted mean distortion for T frames.
(x) T_RM_*d*: root-mean distortion for severity class *d* of T frames.
(xi) T_RM: root-mean distortion for T frames.
(xii) REF_0: the loudness of the lower 3.5 subbands of the reference signal.
(xiii) REF_1: the loudness of the upper 3.5 subbands of the reference signal.

B. GLOBAL MARS MODEL

The basis functions BF*n*, where *n* is an integer, and the regression equation of the global model are listed below:

BF3 = max(0, I_P − 0.270);
BF4 = max(0, 0.270 − I_P);
BF6 = max(0, 33.581 − REF_1);
BF8 = max(0, 0.725 − V_B_2);
BF10 = max(0, 0.131 − I_B_0);
BF12 = max(0, 1.731 − V_B_2_2);
BF13 = max(0, V_P_2 − 0.710);
BF17 = max(0, I_WM_1 − 0.177);
BF20 = max(0, 0.758 − V_P_VUV);
BF23 = max(0, V_P − 0.422);
BF24 = max(0, 0.422 − V_P);
BF25 = max(0, V_O_0 − 2.284);
BF28 = max(0, 0.031 − U_O_6_1);
BF30 = max(0, 0.134 − I_B_5);
BF41 = max(0, V_RM − 0.786);
BF42 = max(0, 0.786 − V_RM);
BF44 = max(0, 0.070 − I_B_1_0);
BF50 = max(0, 0.390 − U_O_4);
BF52 = max(0, 1.657 − U_B_4);
BF62 = max(0, 0.132 − U_O_5);
BF68 = max(0, 0.331 − V_P_1);
BF75 = max(0, V_B_0_1 − 0.337061E-08);
BF154 = max(0, V_O_0_2 − 2.036);
BF169 = max(0, V_O_5 − 0.548);

$$\begin{aligned}
\text{Objective MOS}& \\
&= 2.534 + 6.738 \ast \text{BF3} - 1.833 \ast \text{BF4} \\
&\quad - 0.040 \ast \text{BF6} - 1.331 \ast \text{BF8} - 2.616 \ast \text{BF10} \\
&\quad + 0.600 \ast \text{BF12} + 1.981 \ast \text{BF13} + 4.820 \ast \text{BF17} \\
&\quad + 3.847 \ast \text{BF20} + 3.481 \ast \text{BF23} - 6.184 \ast \text{BF24} \\
&\quad - 0.629 \ast \text{BF25} - 5.552 \ast \text{BF28} + 2.977 \ast \text{BF30} \\
&\quad - 1.296 \ast \text{BF41} + 2.655 \ast \text{BF42} - 3.328 \ast \text{BF44} \\
&\quad + 1.833 \ast \text{BF50} - 0.320 \ast \text{BF52} - 4.596 \ast \text{BF62} \\
&\quad - 1.257 \ast \text{BF68} - 0.476 \ast \text{BF75} + 0.577 \ast \text{BF154} \\
&\quad + 1.585 \ast \text{BF169}.
\end{aligned}$$

ACKNOWLEDGMENTS

We thank Nortel Networks for their financial support. We also thank the reviewers for helping to substantially improve the presentation in this paper.

REFERENCES

[1] D. G. Jamieson, V. Parsa, M. Price, and J. Till, "Interaction of speech coders and atypical speech, II: effects on speech quality," *Journal of Speech Language & Hearing Research*, vol. 45, pp. 689–699, 2002.

[2] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Commun. Mag.*, vol. 26, no. 10, pp. 36–44, 1988.

[3] A. E. Conway, "A passive method for monitoring voice-over-IP call quality with ITU-T objective speech quality measurement methods," in *Proc. IEEE International Conference on Communications (ICC '02)*, vol. 4, pp. 2583–2586, New York, NY, USA, April–May 2002.

[4] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union, Geneva, Switzerland, August 1996.

[5] ITU-T Rec. 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," International Telecommunication Union, Geneva, Switzerland, February 2001.

[6] ITU-T Rec. P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunication Union, Geneva, Switzerland, May 2004.

[7] R. F. Kubichek, D. Atkinson, and A. Webster, "Advances in objective voice quality assessment," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM '91)*, vol. 3, pp. 1765–1770, Phoenix, Ariz, USA, December 1991.

[8] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, New York, NY, USA, 2nd edition, 1990.

[9] S. Voran, "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 4, pp. 371–382, 1999.

[10] S. Voran, "Objective estimation of perceived speech quality. II. Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 4, pp. 383–390, 1999.

[11] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[12] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1984.

[13] J. E. Schroeder and R. F. Kubichek, "L_1 and L_2 normed cepstral distance controlled distortion performance," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM '91)*, vol. 1, pp. 41–44, Victoria, BC, Canada, May 1991.

[14] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, The Netherlands, 1995.

[15] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '01)*, vol. 2, pp. 749–752, Salt Lake City, Utah, USA, May 2001.

[17] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *IEE Proceedings of Vision, Image and Signal Processing*, vol. 141, no. 3, pp. 203–208, 1994.

[18] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Workshop on Speech Coding Proceedings*, pp. 144–146, Porvoo, Finland, June 1999.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, CRC Press, Boca Raton, Fla, USA, 1984.

[20] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.

[21] N. Suzuki, S. Kirihara, A. Ootaki, M. Kitajima, and S. Nakamura, "Statistical process analysis of medical incidents," *Asian Journal on Quality*, vol. 2, no. 2, pp. 127–135, 2001.

[22] K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olshen, and K. L. Oehler, "Bayes risk vector quantization with posterior estimation for image compression and classification," *IEEE Trans. Image Processing*, vol. 5, no. 2, pp. 347–360, 1996.

[23] P. Sephton, "Forecasting recession: can we do better on MARS?" *Federal Reserve Bank of St. Louis Review*, vol. 83, no. 2, pp. 39–49, 2001.

[24] T. Ekman and G. Kubin, "Nonlinear prediction of mobile radio channels: measurements and MARS model designs," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, vol. 5, pp. 2667–2670, Phoenix, Ariz, USA, March 1999.

[25] ITU-T Rec. G.729 - Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," International Telecommunication Union, Geneva, Switzerland, November 1996.

[26] ETSI EN 301 708 V7.1.1, "Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels," Euro. Telecom. Stds. Inst., December 1999.

[27] R. F. Kubichek, E. A. Quincy, and K. L. Kiser, "Speech quality assessment using expert pattern recognition techniques," in *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM '91)*, pp. 208–211, Victoria, BC, Canada, June 1989.

[28] S. Voran, "Advances in objective estimation of received speech quality," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Porvoo, Finland, June 1999.

[29] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech, and Audio Processing*, vol. 1, no. 1, pp. 3–14, 1993.

[30] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[31] L. Chistovich and V. V. Lublinskaya, "The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli," *Hearing Research*, vol. 1, no. 3, pp. 185–195, 1979.

[32] S. Voran, "A simplified version of the ITU algorithm for objective measurement of speech codec quality," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 1, pp. 537–540, Seattle, Wash, USA, May 1998.

[33] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," International Telecommunication Union, Geneva, Switzerland, February 1998.

[34] A. W. Rix, "A new PESQ scale to assist comparison between P.862 PESQ score and subjective MOS," ITU-T SG12 COM12-D86, May 2002.

[35] A. Abraham, "Analysis of hybrid soft and hard computing techniques for forex monitoring systems," in *Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '02)*, vol. 2, pp. 1616–1622, Honolulu, Hawaii, USA, May 2002.

[36] M. Stone, "Cross-validation choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B*, vol. 36, pp. 111–147, 1974.

Wei Zha received his B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, both in electronics engineering. He worked in the Department of Electronics Engineering, Shanghai Jiao Tong University, Shanghai, China. He received his Ph.D. degree in electrical and computer engineering from Queen's University, Kingston, Ontario, Canada, in 2002. From 2002 to 2003, he worked on speech quality measurement at Queen's University. From 2003 till the end of 2004, he was with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, holding an NSERC Fellowship. Since January 2005, he has been with Schlumberger Houston Tech Center.

Wai-Yip Chan (usually known as Geoffrey Chan) received his B.Eng. and M.Eng. degrees from Carleton University, Ottawa, Canada, and his Ph.D. degree from the University of California at Santa Barbara, all in electrical engineering. He is currently an Associate Professor of electrical and computer engineering at Queen's University, Kingston, Canada. Previously, he was on the faculty of Illinois Institute of Technology, Chicago, and McGill University, Montreal. He also worked at the Communications Research Centre and Bell Northern Research (now Nortel Networks), Ottawa, where he acquired industrial experience ranging from embedded DSP algorithm to VLSI circuit design for speech processing. His current research interests are in the area of multimedia signal compression and communications. He served as a Technical Program Cochair of the 2000 IEEE Workshop on Speech Coding, and received a CAREER award from the US National Science Foundation.