A Data Mining Approach to Objective Speech Quality Measurement

Wei Zha and Wai-Yip Chan
Department of Electrical and Computer Engineering, Queen's University
Kingston, Ontario, Canada K7L 3N6
{wzha, chan}@ee.queensu.ca

Abstract -
Existing objective speech quality measurement algorithms still fall short of the measurement accuracy that can be obtained from subjective listening tests. We propose an approach that uses statistical data mining techniques to improve the accuracy of auditory-model based quality measurement algorithms. We present the design of a novel measurement algorithm using the multivariate adaptive regression splines (MARS) method. A large set of speech distortion features is first created. MARS is used to find a small set of features that provide the best estimate ("model") of speech quality. One appeal of the approach is that the model size can scale with the amount of speech data available for learning. In our simulations, the new algorithm furnishes significant performance improvement over PESQ.

I. Introduction

Speech quality measurement is an important problem for the telecommunication networks operators, especially in recent and future years when voice can be transmitted in tandem through different sub-networks, such as wireline, wireless and IP networks. An accurate, robust, and low-complexity speech quality estimation method that can be used in a variety of network conditions would be very useful.

The most reliable method of measuring speech quality today is subjective listening tests. A group of listeners are asked to score the speech they hear according to the absolute categorical rating (ACR) scale. The average of these scores, the subjective mean opinion score (MOS), is widely used to characterize the performance of speech codecs and transmission equipment and networks. Subjective tests are time consuming and costly. In contrast, objective methods can be implemented by computers and embedded into network nodes for real-time speech quality monitoring and control.

In this paper, we propose a new objective quality measurement method. The human speech quality judgment process can be divided into two parts. The first part is the conversion of the received speech signal into auditory nerve excitations for the brain. This part is well documented as auditory periphery system models in the literature. The second part is cognitive processing in the brain, where compact features related to anomalies in the speech signal are extracted and integrated to a final speech quality. This part is largely unknown and is difficult to emulate. Existing objective speech quality methods attempt to approximate this second part, often heuristically. In this paper, we propose a new method based on statistical data mining.

Statistical techniques have been highly successful in advancing the performance of speech recognizers, and likewise can also be exploited to circumvent the need for an anthropomorphic cognitive model. Data mining as a form of machine learning can help us to identify characteristics of speech signals that are well correlated with speech quality. Moreover, data mining provides a means for us to design *scalable* quality estimators. It is highly desirable to have an estimator that can scale with the amount of data available for learning the cognitive mapping. New forms of speech degradations arise as a result of newly collected learning samples, new transmission environments, new speech codecs, etc. Statistical techniques enable us to design best-size estimators for a given amount of learning data, and adapt to new data.

An important step of our proposed method is feature selection and estimator optimization. We first extract a large number of features from the distortion surface (over time-frequency) between the original speech signal and the degraded speech signal. These features are context sensitive, i.e., they are based on local properties. Local properties are determined via segmentation and classification. The features are processed through statistical data mining methods for selection and estimator optimization. In this paper, we focus on one particular data mining method: multivariate adaptive regression splines (MARS) [6].

Our proposed method also has the advantage of simplicity of implementation. The auditory processing model is simplified, in comparison with existing quality estimation algorithms. Furthermore, the computational requirement of feature extraction and quality estimation is only a small fraction of the auditory processing part.

II. Existing Methods

Early speech quality estimators were for estimating the quality of waveform speech coders. The estimators rely on the difference between the clean speech waveform and the coded (degraded) speech waveform, effecting a waveform matching criterion. Representative estimators include the signal-to-noise ratio (SNR) and segmented SNR. Low-bit-rate speech coders do not necessarily preserve the original waveform, so that waveform matching is not appropriate. Speech quality measurement algorithms based on auditory models do not require waveform matching. Algorithms of this type include BSD (Bark spectral distortion) [1], MNB (measuring normalizing block) [2], PSQM [3], and PESQ (perceptual evaluation of speech quality) [4]. BSD was the first to use a precise human

auditory model for speech quality measurement, while ITU-T standard P.862 PESQ offers the current "state-of-the-art" performance.

A major difference among the above auditory-model based methods is in the processing of the auditory error surface. MNB uses a hierarchical structure of integration, over different time and frequency interval lengths. PESQ uses a three step integration, 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 integration performed in the three steps. The integrations are *ad hoc* in nature and not based on cognitive insight [5].

III. Proposed Method

The proposed method consists of two main blocks: auditory processing and cognitive mapping.

A. Auditory processing

Human auditory processing is approximated by the following processing steps. The speech signal is first divided into overlapping frames. The spectral power density of each frame is obtained using FFT. Hertz-to-Bark frequency transformation is performed by summing an appropriate set of power density coefficients. The summed powers are then converted to subjective loudness using Zwicher's law [9]. The final frequency decomposed signal for each speech frame is in sone/Bark unit. In our method, the signal is decomposed into to 7 subbands, with each subband roughly 2.5 Bark wide, for telephone bandwidth speech.

B. Cognitive Mapping - Feature Extraction

The design of the "cognitive" mapping in our proposed scheme is shown in Fig. 1. We first extract a large number of features from the auditory processed clean speech signal and degraded speech signal.

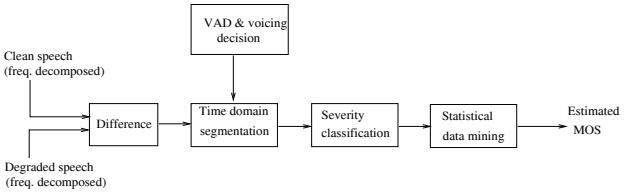


Fig. 1. Cognitive mapping design

The clean and degraded speech signals, decomposed into subjective loudness distributions over bark frequency and time, are first subtracted to form the difference. The difference over the entire speech file corresponds to a distortion surface over time-frequency. The goal of the cognitive mapping is to integrate the distortion surface by segmentation, classification, and simple integration.

The frequency decomposed 7-subband distortions for each frame are classified by a two-stage process. The first stage is time domain segmentation based on voice activity detection (VAD) and voicing decisions. Each speech frame is classified into one of the three categories: inactive, voiced, or unvoiced. As a result, the distortion in each time-frequency bin gets classified into one of $3 \times 7 = 21$ classes.

The distortions from the first stage are further classified by the severity of the frame-distortion, into three different categories, small, medium, and large, using simple thresholding. After two stages of classification, the distortions can be assigned to one of $3 \times 21 = 63$ classes.

The distortions in each of the 63 classes are averaged using L_2 norm. The integrated distortion from each class is called a *feature*. Other types of features calculated include rank-ordered distortions, weighted mean distortion, probability of each type of speech frames. A total of 209 features are available for data mining.

C. Cognitive Mapping - Data Mining Using MARS

MARS [6] builds large regression models over two processing steps. The "forward" step recursively partitions the data domain into smaller regions. In each recursion step, a feature variable is selected for partitioning perpendicular to the variable. Two spline "basis functions," one for each of the two newly created partition regions, are added to the model under construction. The feature variable to choose and the point of partition are found by brute-force search. An overly large model is built initially. In the second "backward" step, basis functions that contribute least to performance are deleted one by one. MARS has been used to forecast recession [7], predict customer spending, and predict radio-channel power [8].

From the large number of features extracted from the distortion surface, MARS is used to find a small subset of features to form the speech quality estimator. The subset of feature variables, together with the particular manner of combining them, are jointly optimized to produce the most statistically consistent estimate of subjective MOS. The estimator, or "data model," is optimized to the "right" model size, for a given amount of subjectively scored speech training data, using statistical validation techniques.

We note that the *statistical data mining* block in Fig. 1 is for the design phase only. Once the feature selection and combining are optimized in the design phase, the block is replaced by a simple mapping block.

IV. Experiment Results

We compare our proposed method to the current state-of-the-art, the ITU-T P.862 standard PESQ. The speech databases used in our experiments include a mixed wireline/wireless database, two wireless databases (IS-96A and IS-127 EVRC), and seven coded speech databases (English, French, Japanese and Italian) in ITU-T P-Series Supplement 23. We combine these ten databases into a global database comprising 1760 speech file pairs. 90% of the sentence pairs in the global database are randomly assigned to a training set and the rest to a test set.

Performance is measured by the correlation of the predicted quality score to subjective MOS, and also by the root mean square of the prediction residue (RMSE) after regression to subjective MOS. Both the correlation and RMSE are calculated using per-condition averaged MOS, similar to the way PESQ test results have been reported.

Table I shows a series of models as a function of training ratio. Here training ratio is defined as the number of data samples in the training set divided by the "model size," which is the effective number of coefficients in the regression equation. Both the training and test results for the MARS models with different training ratios are shown, where R denotes correlation and $\%$ denotes the percentage of reduction in RMSE relative to PESQ. For PESQ, we optimize a third-order regression polynomial on the training set to obtain $R = 0.8212$ and $RMSE = 0.4597$ on the training set, and $R = 0.7953$ and $RMSE = 0.4689$ on the test set. On our data, PESQ using the optimized polynomial performs somewhat better than the PESQ-LP mapping suggested in [12]. We choose a model that has close performance on both the training and test sets. From Table I, the model with training ratio of 28.8 is chosen for the following experiments.

In Table II, we show the correlation and RMSE when the model with training ratio of 28.8 is applied to each database. We see a range of improvements for individual databases, with the exception of the performance degradation incurred for the Wireless IS-96A database. On the average, 19% reduction in root-mean-square MOS prediction error relative to PESQ is obtained. The scatter plots in Fig. 2 compare the correlation with subjective MOS for ITU-T Sup23 Exp3A speech database.

Let x_i and y_i denote realizations of random variables X and Y, respectively. The correlation R is calculated by Pearson's formula:

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

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

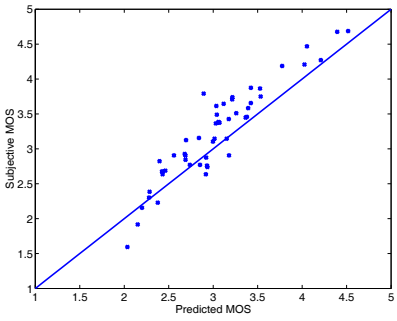
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N}}. \tag{2}$$

Suppose the relationship between X and Y can be modeled as $Y = aX + \epsilon + bias$, where a is a scale factor, ϵ represents zero-mean noise, and $bias$ systematic bias. Then, R and RMSE satisfy the following relationship:

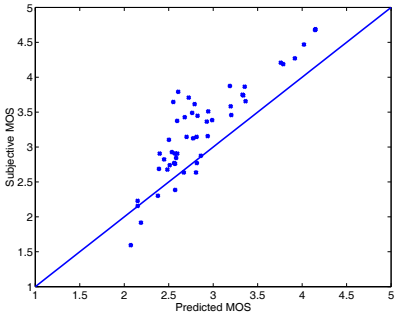
$$\text{RMSE}^2 = \sigma^2(1 - R^2) + bias^2 \tag{3}$$

where σ^2 is the variance of the subjective MOS in a database. The equation states that RMSE is the sum of unexplained variance in the linear regression model, and bias error between subjective MOS and estimated MOS.

We use equation (3) to interpret the relationship between the correlation and RMSE values in Table II. Table II shows



(a) Proposed method



(b) PESQ method

Fig. 2. Scatter plots of estimated MOS versus subjective MOS.

that PESQ incurs large RMSE on databases Exp1D, Exp3A, and Exp3D, even though R is quite high for databases Exp1D and Exp3D. The large RMSE can be attributed to biases between individual databases and the global database. (Note that the correlation calculation excludes the biases between individual databases and the global database.) The MARS model is shown to be able to adapt to individual databases, thus reducing the bias component of the RMSE.

We have compared the results obtained above with another method of finding the best model size, namely n-fold cross validation. The results are similar, and omitted for brevity.

We have also experimented with different database sizes to investigate the scalability of the proposed method. A smaller global database comprising only the seven ITU-T Supplement 23 databases is used. We obtained almost the same RMSE performance gains as presented above. The model size is reduced, and the training ratio changes to 34.8. Thus, the proposed method can be scaled to incorporate new data as it becomes available.

V. Conclusion

A new objective speech quality measurement algorithm designed based on statistical data mining is introduced. In our simulations, our algorithm provides greater measurement accuracy than the PESQ standard algorithm. Our algorithm is also computationally simple to implement and scalable to the amount of data available.

TABLE I

MARS MODEL RESULTS AS A FUNCTION OF TRAINING RATIO

Training ratio	Training			Testing		
	R	RMSE	%	R	RMSE	%
4.66	0.9307	0.2947	35.9	0.8064	0.4644	1.0
5.19	0.9268	0.3026	34.2	0.8232	0.4423	5.7
9.18	0.9068	0.3397	26.1	0.8390	0.4199	10.5
9.46	0.9057	0.3416	25.7	0.8400	0.4185	10.7
16.25	0.8887	0.3694	19.6	0.8519	0.4026	14.1
28.8	0.8749	0.3902	15.1	0.8661	0.3844	18.0
30.17	0.8609	0.4087	11.1	0.8687	0.3923	16.3
39.6	0.8647	0.4047	12.0	0.8564	0.3978	15.2

TABLE II

PERFORMANCE COMPARISON: VARIATION OVER CONDITIONS ONLY

Database	Correlation		RMSE		Percentage Reduction in RMSE (%)
	Proposed Method	PESQ	Proposed Method	PESQ	
ITU-T Sup23 Exp1A (French)	0.9344	0.9360	0.2905	0.3709	21.7
ITU-T Sup23 Exp1D (Japanese)	0.9430	0.9568	0.2377	0.4655	48.9
ITU-T Sup23 Exp1O (English)	0.9690	0.9608	0.2446	0.2857	14.4
ITU-T Sup23 Exp3A (French)	0.9324	0.8833	0.3148	0.4886	35.6
ITU-T Sup23 Exp3C (Italian)	0.9451	0.9533	0.3248	0.3721	12.7
ITU-T Sup23 Exp3D (Japanese)	0.9383	0.9435	0.2968	0.4688	36.7
ITU-T Sup23 Exp3O (English)	0.9379	0.9302	0.2650	0.3452	23.2
Wireless EVRC	0.7968	0.8124	0.2361	0.2427	2.7
Wireless IS-96A	0.6471	0.6209	0.2433	0.2241	-8.6
Mixed	0.9264	0.9243	0.2620	0.2762	5.1
Average					19.24

REFERENCES

[1] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 819-829, Jun. 1992.

[2] S. Voran, "Objective Estimation of Perceived Speech Quality - Part I: Development of the Measuring Normalizing Block Technique," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 371-382, Jul. 1999.

[3] ITU-T Rec. P.861, "Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs," International Telecommunication Union, Geneva, Switzerland, Aug. 1996.

[4] 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, Feb. 2001.

[5] 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 Proc.- Vis. Image Signal Process.*, vol. 141, pp. 203-208, Jun. 1994.

[6] J.H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1-141, March 1991.

[7] P. Sephton, "Forecasting Recession: Can We Do Better on MARS?" *Federal Reserve Bank of St. Louis Review*, vol. 83(2), pp.39-49, Mar./Apr. 2001.

[8] T. Ekman, and G. Kubin, "Nonlinear prediction of mobile radio channels: measurements and MARS model designs," *ICASSP'99*, pp.2667-2670.

[9] E. Zwicker and H. Fastl, "Psychoacoustics - Facts and Models," Springer-Verlag, second edition, 1990.

[10] S. Voran, "A Simplified Version of the ITU Algorithm for Objective Measurement of Speech Code Quality", *ICASSP'98*, pp.537-540.

[11] A.E. Conway, "A Passive Method for Monitoring Voice-over-IP call Quality with ITU-T Objective Speech Quality Measurement Methods," *Proc. Intl. Conf. Commun. 2002*, vol. 4, pp. 2583-2586, Apr.-May 2002.

[12] 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.