Feature Mining for GMM-Based Speech Quality Measurement

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

***Abstract*— We propose a novel approach to objective speech quality measurement using feature mining and Gaussian mixture models (GMMs). A large pool of perceptual distortion features is extracted from the speech signal and data mining techniques are used to sift out the most relevant feature variables from the pool. We examine using multivariate adaptive regression splines (MARS), classification and regression trees (CART), a hybrid CART-MARS scheme, and the sequential forward selection (SFS) algorithm for data mining. For our speech databases, the SFS algorithm provides best performance with a five-feature, three-component GMM. A reduction of 21.7% in root-mean-squared mean opinion score estimation error is obtained in comparison with ITU-T P.862 PESQ.**

I. Introduction

The telecommunications industry is going through a phase of rapid development. New services and technologies emerge continuously. Customers now have the freedom of selecting from as array of telecommunications services at affordable cost. Faced with offering voice services over increasingly heterogenous network connections, the evaluation of speech quality is becoming critically important for the service provider, serving as an instrument for monitoring and improvement of quality of service and network capacity.

Traditionally, the most reliable way to measure the quality of a speech signal was through the use of subjective speech quality assessment tests such as MOS (mean opinion score) tests [1]. These tests are highly unsuitable for online quality measurement and are also very expensive and time consuming. The research described below is motivated by the fact that objective methods have replaced subjective testing, allowing computer programs to automate speech quality measurement in real time, making them suitable for field applications. The International Telecommunications Union ITU-T P.862 standard, also known as Perceptual Evaluation of Speech Quality (PESQ) [2] is the latest objective quality measurement standard algorithm. Nevertheless, the algorithm still falls short of the accuracy that can be obtained from subjective listening tests with sizable listener panels.

In [3] and [4], an approach is introduced that uses data mining techniques to improve the accuracy of auditory-model based quality measurement; significant performance improvement over PESQ was reported. Here we extend the work of [3] and [4] by proposing a novel, simple yet robust, method of speech quality estimation based on Gaussian mixture models (GMMs).

TABLE I
Subjective Rating Scale for the Mean Opinion Score - MOS

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

A large pool of feature measurements is created from the distortion surface between the original speech signal and the degraded speech signal. First, we use four statistical data mining methods, multivariate adaptive regression splines (MARS) [5], classification and regression trees (CART) [6], a hybrid CART-MARS technique, and the sequential forward selection (SFS) algorithm [7] to sift out good features. We then model the joint density of these features ($\mathbf{x}$) with the subjective MOS (y) as a Gaussian mixture. We use this model to derive the minimum mean squared error (MMSE) estimate, $E[y|\mathbf{x}]$, of the subjective MOS value. Simulations show that our approach outperforms PESQ by as much as 21.7% in root-mean-squared MOS estimation error.

II. Objective Speech Quality Estimation

Traditionally, the only way to measure the perceived quality of a speech signal was through the use of subjective testing, i.e, a group of qualified listeners are asked to rate the speech they had just heard according to the scale given in Table I. The average of the listener scores is the subjective MOS. This is the most reliable method of speech quality assessment but it is very expensive and time consuming, making it unsuitable for frequent or rapid applications. Hence, models have been developed to identify audible distortions through an objective process based on human perception. This objective method can be implemented by computer programs and can be used in real time measurement of speech quality.

Several perceptual models for comparing the original and the coded speech signal have been proposed [8]–[11]. The quality of classical coding algorithms is estimated by waveform matching using signal-to-noise ratio (SNR) and the segmental SNR. These are easy to implement, have straightforward interpretations and can estimate the quality of speech in waveform-preserving systems. But newer generation speech coders do not preserve the waveform so these measures are of little relevance.

Researchers have employed models of human auditory perception in their estimation of perceived speech quality. It is known that the peripheral auditory system of human preprocesses information and "compact" feature extraction is done in higher-level brain functions. The human decision is based on this compacted data. An adequate model should emulate this biological preprocessing and higher-level functions, and deliver ratings that have high correlations with the subjective results. The preprocessing part is relatively well understood but the higher-level brain functions are difficult to model.

A. GMMs for Speech Quality Estimation

Gaussian mixture models have been used extensively in speech processing, especially in speech recognition. The reasons behind this widespread use are not coincidental: (1) univariate Gaussian densities have a simple and concise representation, depending uniquely on two parameters, mean and variance, and (2) the Gaussian mixture distribution is universally studied and its behaviors are widely-known [12].

At a cost of extra parameters, GMMs improve on Gaussians by allowing asymmetry and multimodality. In principle, GMM can approximate any probability density function to an arbitrary accuracy. Let $\mathbf{u}$ be an K-dimensional vector, a Gaussian mixture density is a weighted sum of M component densities

$$p(\mathbf{u}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}) = \sum_{i=1}^{M} \alpha_i . b_i(\mathbf{u}) \tag{1}$$

where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u})$, $i = 1, ..., M$ are the K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$.

GMMs can assume several different forms, depending on the type of covariance matrices. The two most widely used are full and diagonal covariance matrices. If K is the dimension of the feature vector and M the number of Gaussian components, then the number of parameters that have to be estimated during training is given by $\frac{M}{2}(K^2 + 3K + 2)$ for full matrices and $M(2K+1)$ for diagonal matrices. The effect of using M full covariance matrices can be equivalently obtained by using a larger set of diagonal covariance Gaussians [13].

The GMM for speech quality estimation is built on perceptual feature variables. The variables are obtained from mining a large pool of candidate feature variables. These candidate features are obtained by classifying perceptual distortions into a variety of contexts, as proposed in [3].

First, the clean and degraded signals are split into 7 frequency bands. The spectral power distortion between the clean and degraded speech signals is then found. Time segmentation labels the speech frames as "active" or "inactive". Active frames are further classified into voiced or unvoiced. The total distortion of each frame is given severity classifications of "low", "medium", or "high" by simple thresholding. Distortion samples in time-frequency bins are thus labelled according to its frequency band, time-segmentation type, and severity level.

Additional contexts are created where each subband is further labelled with the rank order obtained by ranking the 7 distortions in a frame in the order of decreasing magnitude. Weighted mean and root-mean distortions, probability of each frame type and the lowest-frequency band and highest-frequency band energy of the clean speech frames are also used to form a pool of 209 candidate features.

Statistical data mining is used to sift out the most relevant variables from the pool of variables. The top-5 most important feature variables as ranked by MARS, CART, a CART-MARS hybrid configuration, or the SFS algorithm will be used here. We model the joint density of the top-5 most important feature variables ($\mathbf{x}$) with the subjective MOS (y) as a Gaussian mixture given by (1) with $\mathbf{u} = [y, \mathbf{x}]^T$. We then predict the value of the subjective MOS, y, given the observed values of the 5-dimensional feature vector, $\mathbf{x}$. The MMSE estimate of y given $\mathbf{x}$, namely $E[y|\mathbf{x}]$, is [14]

$$E[y|\mathbf{x}] = \sum_{i=1}^{M} h_i(\mathbf{x})[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x} - \boldsymbol{\mu}_j^x)] \tag{2}$$

where $h_i(\mathbf{x})$ denotes the probability that the i^{th} Gaussian component of the marginal predictor density $p(\mathbf{x})$ generated the vector $\mathbf{x}$ and is given by

$$h_i(\mathbf{x}) = \frac{\frac{\alpha_i}{|\boldsymbol{\Sigma}_i^{xx}|^{1/2}} \; e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i^x)^T(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x}-\boldsymbol{\mu}_i^x)\right)}}{\sum_{k=1}^{M} \frac{\alpha_k}{|\boldsymbol{\Sigma}_k^{xx}|^{1/2}} \; e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k^x)^T(\boldsymbol{\Sigma}_k^{xx})^{-1}(\mathbf{x}-\boldsymbol{\mu}_k^x)\right)}}. \tag{3}$$

The covariance matrix of the i^{th} GMM component is

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \Sigma_i^{yy} & \boldsymbol{\Sigma}_i^{yx} \\ \boldsymbol{\Sigma}_i^{xy} & \boldsymbol{\Sigma}_i^{xx} \end{pmatrix}.$$

If the covariance matrices are restricted to be diagonal, the least squares estimate simplifies to

$$E[y|\mathbf{x}] = \sum_{i=1}^{M} h_i(\mathbf{x}) \boldsymbol{\mu}_i^y. \tag{4}$$

This restriction has to be used with care, as it can result in large estimation errors when there exists a significant amount of correlation between the predictor and the response variables, i.e. $\boldsymbol{\Sigma}_i^{yx}$ are far from zero.

III. Experimental Results

We compare our algorithm to PESQ using MOS labelled speech databases. The performance of the algorithms is assessed by the correlation between subjective MOS w_i and objective MOS y_i, using Pearson's formula

$$R = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{5}$$

where $\bar{w}$ is the average of w_i, and $\bar{y}$ is the average of y_i. MOS measurement accuracy is assessed using the root-mean-square MOS error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(w_i - y_i)^2}{N}}. \tag{6}$$

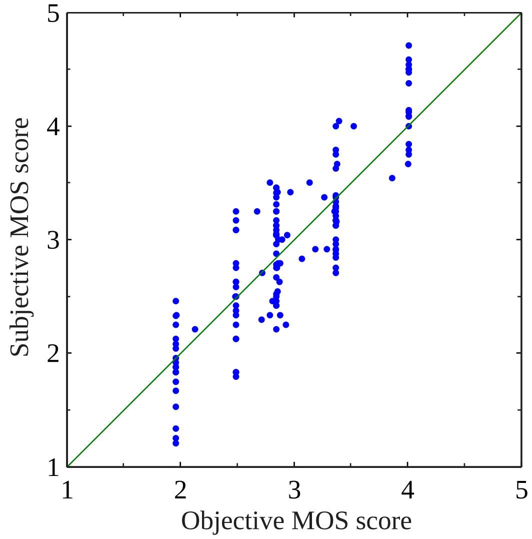

Fig. 1. Subjective MOS *versus* Objective MOS for CART selected features using five diagonal Gaussian components.

In [3], RMSE is shown to be the sum of unexplained variance in the regression model and the bias error between subjective MOS and objective MOS. The calculation of R does not take into consideration this bias error; therefore, unless the estimates are unbiased or all suffer from the same bias, RMSE is a more realistic measure of estimator performance.

The speech databases include seven multilingual databases in ITU-T P-series Supplement 23, two wireless databases and a mixed wireline-wireless database. We combine these ten databases into a global database and then use 10-fold cross validation to measure performance. The global database is randomly divided into 10 data sets of almost equal size. Training and testing is performed in 10 trials, where, in each trial, one of the data sets serves as a test set and the remaining 9 are combined to serve as a training set. Each data set serves as a test set only once. The ten resulting R's and RMSE's are averaged to obtain the cross-validation R and RMSE.

The parameters of the GMM are estimated via the EM algorithm [15]. The algorithm iterations produce a sequence of models with monotonically nondecreasing (log-)likelihood values. Though the EM algorithm converges to a maximum likelihood it has a few drawbacks: it is a greedy algorithm and may converge to a local maximum and not the global maximum. GMMs produced by the EM algorithm are, consequently, sensitive to initialization. We use the *k-means* algorithm [16] to initialize the GMM parameters.

Initial tests show that by using diagonal covariance matrices only a modest improvement over PESQ is achieved. We attribute this to the fact that the features selected by CART and/or MARS have significant correlation amongst them and the use of a small number of diagonal Gaussian components does not compensate for this. When looking at the graph of the subjective MOS *versus* objective MOS for one of the cross-validation trials, we see the penalty of using diagonal matrices (vide Fig. 1). The prominent vertical alignment of points suggests poor prediction performance. The alignment disappears and better prediction performance is obtained when full covariance matrix is used, as we show below.

TABLE II
PERFORMANCE COMPARISON FOR CART SELECTED VARIABLES - FULL COVARIANCE MATRICES

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
GMM-2	0.8569	4.69	0.3892	15.39
GMM-3	0.8611	5.20	0.3860	16.09

TABLE III
PERFORMANCE COMPARISON FOR MARS SELECTED VARIABLES - FULL COVARIANCE MATRICES

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
GMM-2	0.8683	6.10	0.3773	17.52
GMM-3	0.8780	6.35	0.3783	17.98

With full covariance matrices the number of parameters that need to be estimated scales quadratically with the feature space dimension. When dealing with limited data, as in our case, severe problems arise due to singularities and local maxima in the log-likelihood function. Many regularization schemes have been proposed to improve the smoothness and generalization properties of the estimated density function. Here we avert ill-conditioning by adding a small diagonal matrix, namely $\epsilon I_{n\times n}$, to each covariance matrix in each M-step iteration of the EM algorithm. Typically, the optimal value for ϵ is not known a priori. A simple procedure used here is to vary ϵ over a range of values and choose the value that leads to the best performance on the validation data set. We varied ϵ from 0.000001 to 1 and the value that led to best performance was $\epsilon = 0.001$.

The performance results for the feature variables selected by CART and MARS are shown in Tables II and III, respectively. GMM-i stands for a Gaussian mixture model with i components and % indicates percentage increase in R or percentage reduction of RMSE relative to PESQ. The result for PESQ is based on using a 3rd order regression polynomial trained on the global database. On our data, this method outperformed the PESQ-LQ mapping proposed in [17]. The 5 most salient feature variables are listed in Table IV for all four data mining techniques. The variables are defined in the Appendix.

With the correlation between features properly modelled, an average improvement of 4.95% and 15.74% in R and RMSE, respectively, is achieved for CART selected features. Further improvement can be seen for MARS selected features; an average improvement of 6.23% and 17.75% in R and RMSE is achieved.

As can be seen, only voiced frames are captured by the features selected using CART. In whispered speech, all normally voiced phonemes are not vocalized, i.e. they become unvoiced. Such situation, though rare, would cause problems for these features. It can be inferred that the estimator is not sensitive to degradation in the unvoiced regions and the non-

TABLE IV
FEATURE VARIABLES

Rank	MARS	CART	CART-MARS	SFS
1	I_P_VUV	V_WM	I_P_VUV	V_O_1
2	V_B_5	V_O_2	REF_1	I_P_VUV
3	V_B_2	V_O_1	V_WM_2	REF_1
4	V_B_2_2	V_RM	V_B_5	V_B_2
5	U_P_VUV	V_O_0	U_P	V_O_5

TABLE V
PERFORMANCE COMPARISON FOR CART-MARS SELECTED VARIABLES - FULL COVARIANCE MATRICES

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
Full GMM-2	0.8662	5.83	0.3766	18.13
Full GMM-3	0.8734	6.71	0.3724	19.04

TABLE VI
SUBSET OF THE VALIDATION SET

#	MOS	$\hat{MOS}$	$x1$	$x2$	$x3$	$x4$	$x5$
1	3.5	3.843	0.312	0.421	0.662	0.517	1.072
2	3.5	3.598	0.328	0.471	0.703	0.521	1.011
3	3.5	3.037	0.666	0.885	1.194	0.918	1.680
4	3.5	2.309	0.766	1.091	1.538	1.169	2.226
5	3.5	3.038	0.334	0.504	0.895	0.753	1.691
6	3.5	2.308	0.902	1.101	1.448	1.222	2.402

TABLE VII
PERFORMANCE COMPARISON FOR SFS SELECTED VARIABLES - FULL COVARIANCE MATRICES

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
Full GMM-2	0.8834	7.93	0.3649	20.67
Full GMM-3	0.8840	8.00	0.3602	21.70

speech regions. The CART-MARS hybrid scheme uses CART to pre-screen features from the feature candidate pool. The features selected by CART are then used as a smaller feature candidate pool for MARS to sort through. By doing this, features can be drawn from voiced and unvoiced frames. An average improvement of 6.27% and 18.6% in R and RMSE is achieved. The performance results for the feature variables selected by this hybrid scheme are shown in Table V.

Careful analysis of the features selected by the three aforementioned data mining schemes led us to believe that further improvement is possible. A certain trend was noted within the feature variables as shown in Table VI. The table is composed of the subjective MOS, five MARS-selected features ($x1$ to $x5$), and objective MOS ($\hat{MOS}$) estimated using three Gaussian components for six distinct test speech signals. Note that all six test signals have the same subjective MOS, i.e., they have been rated as having, on average, the same objectionable level of distortion. It can be inferred that these six speech signals belong to the same "distortion class" and their feature values should not vary considerably.

Table VI shows that the features selected by MARS, on the contrary, vary considerably and this variation is reflected on the estimates. For test vector 2 we obtain an estimate of 3.598, i.e. an error of 2.8%, but for test vector 6 we obtain an estimate of 2.308, an error of 34%! We conjecture that in order to obtain further improvement better features would have to be used, preferably features that do not vary considerably within the same distortion class. To this end, we use the SFS algorithm. The algorithm starts with the variable that is most correlated with the target variable, and at each step adds a new variable that, together with the previous ones, most accurately predicts the target. A partial F-test is incorporated in the algorithm such that the variables chosen have small variances within each distortion class.

An improvement of 8% and 21.7% in R and RMSE is provided by the SFS algorithm. The performance results of this scheme are shown in Table VII. Table IV shows that the features selected by the SFS algorithm are gleaned from the top three features selected by MARS, CART, and the CART-MARS hybrid scheme.

Figure 2 depicts a scatter plot of the subjective MOS *versus* objective MOS for CART-MARS selected features, using a GMM with three Gaussian components and full covariance matrices. The data shown are for one of the cross-validation trials. We see that the points are no longer aligned with the vertical axis as in the case of diagonal covariance matrices.

A further method for measuring model performance is to plot the distribution of absolute residual errors between objective and subjective MOS [18]. Figure 3 plots the distribution of errors for SFS selected features for one of the trials. As can be seen, almost 78% of the GMM estimates are within 0.50 unit of the subjective MOS.

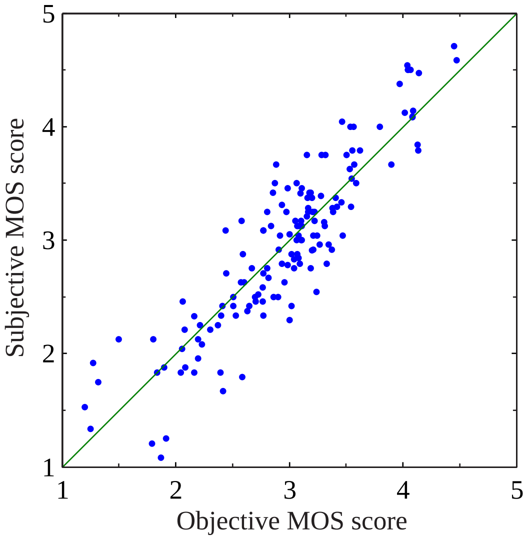

Fig. 2. Subjective MOS *versus* Objective MOS for CART-MARS selected features using three full Gaussian components.

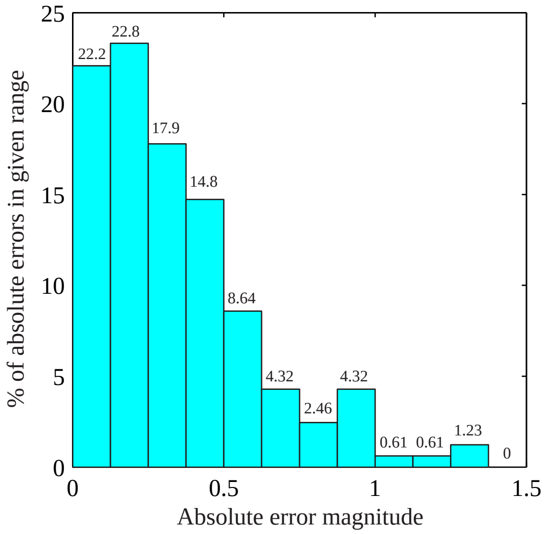

Fig. 3. Residual error distribution for SFS selected features

IV. Conclusion and Further Investigation

A novel objective speech quality estimation algorithm is proposed based on Gaussian mixture modeling. We have investigated the usefulness of features selected by CART, MARS, a CART-MARS hybrid scheme, and the SFS algorithm. The SFS algorithm provided the best performance. The CART-MARS hybrid scheme improved on CART by including features from unvoiced frames. When using diagonal Gaussian components we observed that our approach provided only modest improvement over PESQ. This was attributed to the fact that the five most salient feature variables selected by the data mining techniques were correlated and the use of only five diagonal components was not sufficient to compensate for this. With full Gaussian components we have obtained an improvement over PESQ of 8% and 21.7% in R and RMSE, respectively.

While our results show that feature mining in conjunction with GMM modelling can produce simple estimators that outperform PESQ, the robustness of the estimators is also an important issue. Our use of cross-validation to measure performance offers some robustness. We are currently pursuing other avenues including the choice of feature variables. Ongoing research examines feature selection that directly optimizes GMM estimation performance. Currently, CART/MARS selected features could be suboptimal for GMM estimation. Deeper insights would shed light on why the approach works well or not so well, and whether there is a gap relative to best possible performance.

V. Appendix

Here we describe the feature variables shown in Table IV. The seven subbands are ordered from 0 to 6 and the three distortion severity classes from 0 to 2.

- I_P_VUV: Ratio of the number of inactive frames to the total number of active speech frames;
- U_P_VUV: Ratio of the number of unvoiced frames to the total number of active speech frames;
- U_P: Percentage of unvoiced frames in the speech files;
- REF_1: High-frequency spectral energy of reference signal;
- V_B_i: Distortion for subband *i* of voiced frames, without distortion severity classification;
- V_B_i_j: Distortion for severity class *j* of subband *i* of voiced frames;
- V_O_i: Distortion for ordered subband *i* of voiced frames, without distortion severity classification;
- V_WM_i: Weighted mean distortion for severity class *i* of voiced speech frames;
- V_WM: Weighted mean distortion of voiced speech frames;
- V_RM: Root-mean distortion of voiced speech frames.

References

[1] ITU-T Rec. P.830, “Subjective performance assessment of telephone-band and wideband digital codecs,” International Telecommunication Union, Geneva, Switzerland, Feb. 1996.

[2] 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.

[3] W. Zha and W.-Y. Chan, “A data mining approach to objective speech quality measurement,” in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing 2004*, vol. 1, May 2004, pp. 461–464.

[4] ——, “Voice quality assessment using classification trees,” in *Proc. of the* 37^{th} *Asilomar Conference on Signals, Systems and Computers*, vol. 1, November 2003, pp. 537–541.

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

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks, 1984.

[7] A. Jain and D. Jongker, “Feature selection: evaluation, application, and small sample performance,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, February 1997, pp. 153–158.

[8] S. Wang, A. Sekey, and A. Gersho, “An objective measure for predicting subjective quality of speech coders,” *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, June 1992.

[9] R. Kubichek, E. Quincy, and K. Kiser, “Speech quality assessment using expert pattern recognition techniques,” in *Proceedings of the IEEE Conference on Communications, Computers and Signal Processing*, June 1989, pp. 208–211.

[10] S. Voran, “Objective estimation of perceived speech quality - part I: Development of the measuring normalizing block technique,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4, pp. 371–382, July 1999.

[11] ——, “Objective estimation of perceived speech quality - part II: Evaluation of the measuring normalizing block technique,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4, pp. 383–390, July 1999.

[12] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, “Gaussian mixture density modeling, decomposition and applications,” *IEEE Transactions on Image Processing*, vol. 5, no. 9, pp. 1293–1302, September 1996.

[13] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[14] Z. Ghahramani and M. I. Jordan, “Supervised learning from incomplete data via an EM approach,” in *Advances in Neural Information Processing Systems*, vol. 6. Morgan Kaufmann Publishers, Inc.

[15] A. Dempster, N. Lair, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[16] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[17] 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.

[18] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “PESQ - the new ITU standard for end-to-end speech quality assessment,” in 109^{th} *AES Convention*, no. pre-print 5260, September, pp. 1–18.