Speech Quality Estimation Using Gaussian Mixture Models

Tiago H. Falk[1], *Wai-Yip Chan*[1], *and Peter Kabal*[2]

Department of Electrical and Computer Engineering
[1]Queen's University, Kingston, ON, Canada K7L 3N6
[2]McGill University, Montreal, QC, Canada H3A 2A7
{falkt,chan}@ee.queensu.ca, kabal@ece.mcgill.ca

Abstract

We propose a novel method to estimate the quality of coded speech signals. The joint probability distribution of the subjective mean opinion score (MOS) and perceptual distortion feature variables is modelled using a Gaussian mixture density. The feature variables are sifted from a large pool of candidate features using statistical data mining techniques. We study what combinations of features and mixture model configuration are most effective. For our speech database, a five-feature, three-component GMM furnishes approximately 18% lower root-mean-squared MOS estimation error than ITU-T P.862 PESQ, the current best standard algorithm.

1. Introduction

The evaluation of speech quality is of critical importance in today's telephone networks, be it plain old telephone system, wireless, or voice over Internet, mainly because quality is a key determinant of customer satisfaction. Traditionally, the most reliable way to measure the quality of a speech signal was through the use of subjective testing, i.e, a group of qualified listeners are asked to score the speech they just heard on to a scale from 1 to 5, with 1 corresponding to unsatisfactory speech quality with very annoying and objectionable levels of distortion and 5 corresponding to excellent speech quality and imperceptible level of distortion. The average of these scores is the subjective mean opinion score, MOS [1]. This method of speech quality assessment is highly unsuitable for automation of voice connection quality measurement and is also very expensive and time consuming. Due to these reasons, models have been developed to identify audible distortions through an objective process based on human perception. Objective methods can be implemented by computer programs to automate speech quality measurement in real time. 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. In [3], 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.

In this paper we propose a novel method of speech quality estimation based on Gaussian mixture models (GMMs). First a large pool of feature measurements is created from the distortion surface between the original speech signal and the degraded speech signal. Good features are then selected using two statistical data mining methods, multivariate adaptive regression splines (MARS) [4] and classification and regression trees (CART) [5]. We 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 least squares estimate, $E[y|\mathbf{x}]$, of the subjective MOS value. Simulations show that our approach outperforms both PESQ and the method proposed in [3].

2. Gaussian Mixture Models

Gaussian mixture models have been used extensively within the speech processing community and will be briefly introduced here for the sake of notation. Let $\mathbf{u}$ be an N-dimensional vector, a Gaussian mixture density is a weighted sum of M component densities

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

where $p_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} p_i = 1$, and $b_i(\mathbf{u})$, $i = 1, ..., M$ are the K-variate Gaussian densities each 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 by $M(2K+1)$ for diagonal matrices. The effect of using M full covariance matrices can be obtained by using a larger set of diagonal covariance Gaussians [6].

In this study both cases are considered and the EM (expectation-maximization) algorithm [7] is used for estimation of the weights, means and covariances of the Gaussian components.

2.1. GMM for Speech Quality Estimation

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.

First, the clean and degraded signals are split into 7 frequency bands, each with a bandwidth of approximately 2.4 Bark. 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.

We use CART and/or MARS to sift out the most relevant variables from the candidate pool. We use the top-5 most important feature variables as ranked by MARS or CART. We model the joint density of these features ($\mathbf{x}$) with the subjective MOS (y) as a Gaussian mixture. The goal is to predict the value of the subjective MOS, y, given the observed values of the 5-dimensional feature vector, $\mathbf{x}$. The least squares estimate of y given $\mathbf{x}$, namely $E[y|\mathbf{x}]$, is [8]

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

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

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

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

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

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.

3. Experimental Results

We compare our algorithm to PESQ and to the method proposed in [3], which will be referred to as SDMA (statistical data mining approach) in the sequel. We use MOS labelled speech databases and the performance of each algorithm is assessed using the correlation (R) between the subjective MOS and the predicted MOS, and the root mean squared MOS error (RMSE).

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 10 times, where, each time, 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. 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 since the likelihood for GMMs is not unimodal the algorithm may converge to a local maximum and not the global maximum. GMMs produced by the EM algorithm are sensitive to initialization and may converge to the boundary of the parameter space where the likelihood is unbounded, leading to meaningless estimates. The *k-means* algorithm is used to initialize the GMM parameters.

The performance results for the feature variables selected by MARS and CART are shown in Tables 1 and 2 respectively. GMM-i stands for a Gaussian mixture model with i components and % shows the performance improvement over PESQ. The 5 most salient feature variables are listed in Table 3 for each of the data mining techniques. The variables are defined in the Appendix.

As can be seen, when using diagonal GMMs, an average of 13.03% improvement in RMSE is achieved, while the improvement in R is more modest. This occurs because some of the features selected by MARS and CART

Table 1: Performance Comparison for MARS selected variables - diagonal covariance matrices

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
GMM-3	0.8086	- 1.21	0.4094	11.01
GMM-4	0.8232	0.57	0.4008	12.86
GMM-5	0.8377	2.34	0.3971	13.67

Table 2: Performance Comparison for CART selected variables - diagonal covariance matrices

	R	%	RMSE	%
PESQ	0.8185	N/A	0.460	N/A
GMM-3	0.8315	1.60	0.4035	12.27
GMM-4	0.8395	2.57	0.3957	13.97
GMM-5	0.8531	4.23	0.3938	14.38

Table 3: Feature Variables

Rank	MARS	CART
1	I_P_VUV	V_WM
2	V_B_5	V_O_2
3	V_B_2	V_O_1
4	V_B_2_2	V_RM
5	U_P_VUV	V_O_0

have significant correlation amongst them. This is illustrated with the use of the correlation color map in Figure 1. This figure represents the correlation between the predictor and the response variables selected by MARS. For CART, the color map is similar and will be omitted for brevity. The use of a small number of diagonal Gaussian components does not compensate for this correlation and full covariance matrices are thus needed in order to predict the residual variation in subjective MOS.

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 limit the spectral dynamic range 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. The simplest procedure, and the one used here is to vary ϵ over a range of values and choose the one that leads to the best performance on the validation set. We varied ϵ from 0.000001 to 1 and the value that led to best performance was $\epsilon = 0.001$.

Tables 4 and 5 show the performance improvements by using full covariance matrices. 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. Figure 2 shows the scatter plot of the subjective MOS *versus* objective MOS for MARS-selected features using three Gaussian components with full covariance matrices. It is also worth mentioning that our results outperform SDMA by as much as 6% in RMSE.

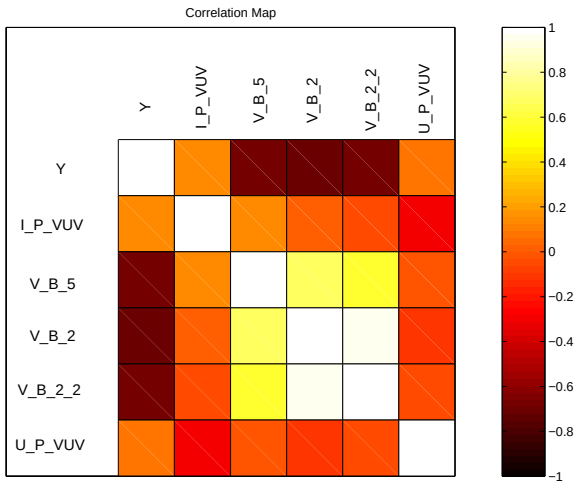

Figure 1: Correlation map for MARS-selected features.

Table 4: 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 5: 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

4. Conclusion

A novel objective speech quality estimation algorithm is proposed based on Gaussian mixture models. When using diagonal Gaussian components we observed that our approach outperforms PESQ in RMSE but the improvement in R is smaller. 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 enough to compensate for this.

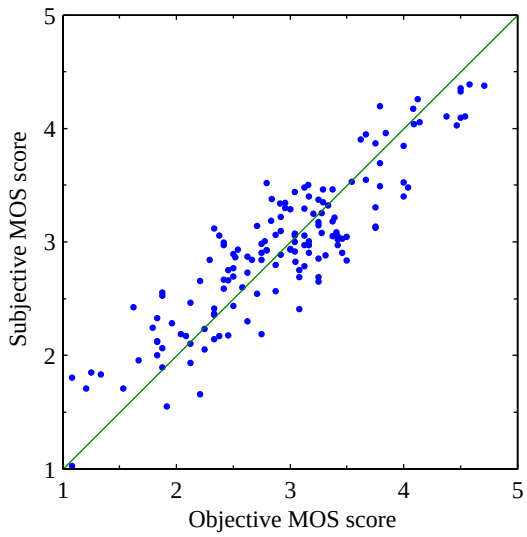

Figure 2: Subjective MOS *versus* Objective MOS for MARS-selected features using three full Gaussian components.

By adding a bias term to the model in [9], RMSE can be shown to be the sum of unexplained variance in the regression model, MOS estimation error due to limited number of listeners (affecting all algorithms equally), 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. So if data is limited, diagonal Gaussian components can be used with an average improvement in RMSE of 13.03%.

In the case a larger dataset is available, one is motivated to use full Gaussian components with which we have obtained an average RMSE improvement over PESQ of 16.75%, or a performance improvement of approximately 6% over SDMA.

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.

5. Appendix

Here we describe the feature variables shown in Table 3. 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;
- 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: Weighted mean distortion of voiced speech frames;
- V_RM: Root-mean distortion of voiced speech frames.

6. 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 Int. Conf. on Acoustics, Speech, and Signal Processing 2004*, vol. 1, May 2004, pp. 461–464.

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

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

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

[7] 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.

[8] 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.

[9] R. Kubichek, D. Atkinson, and A. Webster, "Advances in objective voice quality assessment," in *Proc. of the 1991 Globecom*, pp. 1765–1770.