Objective Speech Quality Assessment Using Gaussian Mixture Models

Tiago H. Falk and Wai-Yip Chan
Department of Electrical and Computer Engineering
Queen's University, Kingston, ON, Canada K7L 3N6
Email: {falkt, chan}@ee.queensu.ca

Abstract

Objective speech quality assessment algorithms provide low-cost and online monitoring of voice calls, replacing costly and time-consuming subjective listening tests. We propose a novel approach to objective speech quality measurements using Gaussian mixture models (GMMs). A large pool of perceptual distortion features is extracted from speech files and multivariate adaptive regression splines (MARS) is used to sift out the most relevant variables from the pool. The five most salient variables are used to construct good GMM estimators of subjective listening quality. Simulation results show that this novel approach outperforms the state-of-the-art objective measurement algorithm, PESQ.

I. Introduction

The evaluation of speech quality is of critical importance in today's telephone networks, be it POTS, wireless or VoIP, mainly because quality is a key determinant of customer satisfaction. Traditionally, the only way to measure the perception of 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 according to a scale from 1 to 5, where 1 corresponds to unsatisfactory speech quality with very annoying and objectionable levels of distortion and 5 corresponds to excellent speech quality and imperceptible level of distortion. The average of these scores is the subjective mean opinion score, MOS. This has been the most reliable method of speech quality assessment but it is highly unsuitable for online monitoring applications and is also very expensive and time consuming. Due to these reasons, models were developed to identify audible distortions through an objective process based on human perception. Objective methods can be implemented by computer programs and can be used in real time monitoring of speech quality. Algorithms for objective measurement of speech quality assessment have been implemented and the International Telecommunications Union has promulgated ITU-T P.862 standard, also known as Perceptual Evaluation of Speech Quality (PESQ), as its state-of-the-art algorithm.

In this paper we propose a novel method of speech quality assessment based on Gaussian mixture models. 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 a statistical data mining method, multivariate adaptive regression splines (MARS) [1]. We model the joint density of these features (x) with the subjective MOS (y) as a Gaussian mixture. We then use this model to derive the least squares estimate, $E[y|x]$, of the subjective MOS value. Simulations show that our approach outperforms PESQ.

II. Gaussian Mixture Models

A Gaussian mixture density is a weighted sum of M component densities

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

where **u** is an N-dimensional vector, $b_i(\mathbf{u})$, $i=1,...,M$ are the mode densities and $p_i\geq 0, i=1,...,M$ are the mixture weights, such that $\sum_{i=1}^{M}p_i=1$. Each mode density is a K-variate Gaussian function with mean vector μ_i and covariance matrix Σ_i. Gaussian mixture densities are, thus, parameterized by three components: the mean vectors, covariance matrices and mixture weights. By varying the number of Gaussians, M, and the three components one can, in principle, approximate **any** complex probability density function, to an arbitrary accuracy.

Since the Gaussian Mixture model uses a discrete set of Gaussian functions, each with their own mean and variance, it is expected that the GMM have several different forms, depending on the shapes of their covariance matrices. The two most widely used forms are full and diagonal covariance matrices. The full covariance matrix is the most powerful Gaussian model, as it fits the data best. The drawback is the fact that it needs lots of data to properly estimate parameters and it becomes costly in high-dimensional feature spaces. 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 $\frac{M}{2}(K^2+3K+2)$.

On the other hand, diagonal covariance matrices are a good compromise between quality and model size. In this case, the total number of parameters that need to be estimated during training is given by $M(2K+1)$. This type is widely used in practice and mainly due to the fact that, since the Gaussian components are acting together to model the overall probability density function, a linear combination of diagonal covariance Gaussians is capable of modelling the correlations between the feature vectors [2].

In this study both cases are considered and the EM (expectation-maximization) algorithm [3] is used for estimation of the weights, means and covariances of the Gaussian components.

A. GMM for Objective Speech Quality Assessment

Motivated by the work of [4] we use an algorithm based on classifying perceptual distortions under a variety of contexts. First, the signals are frequency decomposed into 7 bands. The distortion between the decomposed clean and degraded speech signals is then found. Cognitive mapping is achieved by aggregating cognitively similar distortion events through time segmentation and distortion severity classification. Time segmentation labels the speech frames as “active” or “inactive”. Active frames are further classified into voiced or unvoiced. Distortion severity classification labels the total distortion of each frame as “low”, “medium” or “high” by means of 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 statistical data mining to find underlying patterns or relationships in the data sets and to sift out the most relevant variables from a large pool of candidate variables. We use the top-5 most important feature variables as ranked by MARS. We model the joint density of these features (x) with the subjective MOS (y) as a Gaussian mixture model. The goal is to predict the values of the subjective MOS, *y*, given the observed values of the *5*-dimensional feature vector, *x*. To use GMM’s as regressors, the best least squares estimate of y given x, namely $E[y|x]$, is derived [5]:

$$E[y|x] = \int dy\ y\ p(y|x) = \frac{\int dy\ y\ p(y,x)}{p(x)} = \frac{\int dy\ y\ p(y,x)}{\int dy\ p(y,x)}$$
$$= \sum_{i=1}^{M} h_i(x)[\mu_i^y + \Sigma_i^{yx}\Sigma_i^{xx-1}(x-\mu_j^x)] \qquad (2)$$

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

$$h_i(x) = \frac{\frac{p_i}{|\Sigma_i^{xx}|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_i^x)^T\Sigma_i^{xx-1}(x-\mu_i^x)\right)}{\sum_{k=1}^{M}\frac{p_k}{|\Sigma_k^{xx}|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k^x)^T\Sigma_k^{xx-1}(x-\mu_k^x)\right)}. \qquad (3)$$

The superscripts y and x denote vectors belonging to the response and the predictor variables, respectively. The covariance matrix for the i^{th} component is $\Sigma_i = \begin{pmatrix} \Sigma_i^{yy} & \Sigma_i^{yx} \\ \Sigma_i^{xy} & \Sigma_i^{xx} \end{pmatrix}$.

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

$$E[y|x] = \sum_{i=1}^{M} h_i(x)\mu_i^y. \qquad (4)$$

III. Experimental Results

Here we compare our algorithm to the current state-of-the-art algorithm for voice quality estimation, PESQ, using MOS labelled speech databases. The performance of the algorithm is based on the correlation (R) between the subjective MOS and the predicted MOS. The root mean squared error (RMSE) is used to assess the MOS measurement accuracy.

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. The global database is randomly divided into 10 data sets of almost equal size. Training and testing is thus 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.

As mentioned previously, the parameters of the GMM will be estimated via the EM algorithm. Each EM iteration guarantee’s a monotonic increase in the model’s likelihood (log likelihood) value. It is known that the EM algorithm converges to a maximum likelihood but has a few drawbacks: it is a greedy algorithm and since the likelihood of GMMs are not unimodal it may converge to a local maximum and not the global maximum. This makes the EM algorithm very sensitive to initialization and may converge to the boundary of the parameter space where the likelihood is unbounded, leading to meaningless estimates.

Here the *k-means* algorithm is used to find the initial parameters. It partitions the data into M subsets, each subset populating a region in the feature space. The empirical probability of each subset becomes the initial mixture weights. The mean of the data in each subset becomes the initial mean of the corresponding mixture kernel and the covariance of the data of each subset determines the initial covariance of the respective component.

The performance results for the feature variables selected by MARS are shown in Table I. Diagonal GMM-i stands for a Gaussian mixture model with i components and diagonal covariance matrices. Percentage Increase/Decrease shows the performance improvement over PESQ. As can be seen, small improvement in R is achieved when using diagonal GMMs. On the other hand, an average of 12.31% improvement in RMSE is achieved. This occurs because some of the features selected by MARS have significant correlation amongst them and the use of a small amount of diagonal Gaussian components does not compensate for this. When looking at the graph of the objective MOS *versus* subjective MOS we see the penalty of using diagonal matrices (vide Fig. 1). The prominent vertical alignment of points suggests that full covariance matrices are 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 input 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

TABLE I

PERFORMANCE COMPARISON FOR MARS SELECTED VARIABLES

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

TABLE II

PERFORMANCE COMPARISON FOR MARS SELECTED VARIABLES

	R	Percentage Increase (%)	RMSE	Percentage Decrease
PESQ	0.8185	N/A	0.460	N/A
Full GMM-2	0.8683	6.10	0.3773	17.52
Full GMM-3	0.8780	6.35	0.3783	17.98

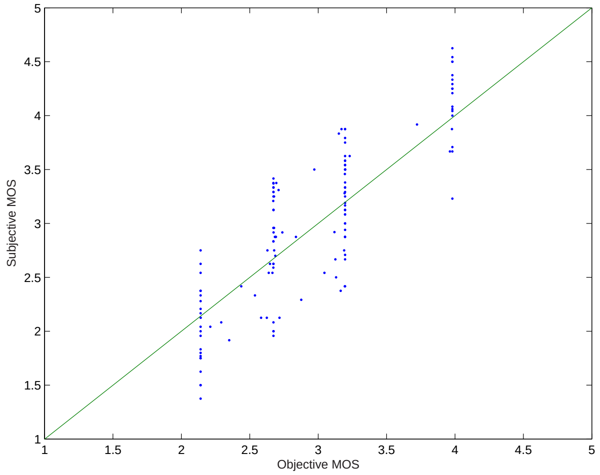

Fig. 1. Objective MOS *versus* Subjective MOS for MARS-selected features using four diagonal Gaussian components.

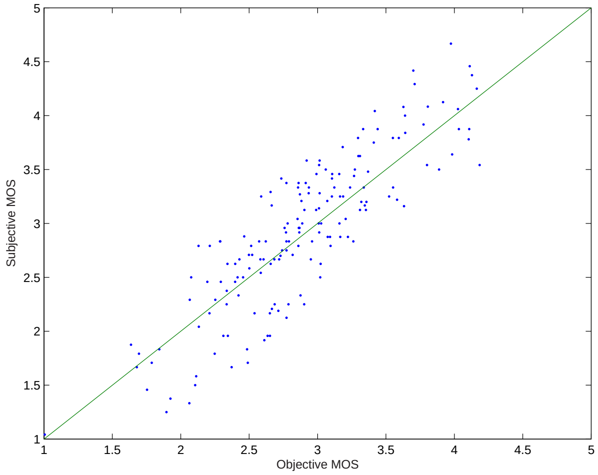

Fig. 2. Objective MOS *versus* Subjective MOS for MARS-selected features using three full Gaussian components.

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 in [6] 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 $\epsilon = 0.001$.

Table II shows the performance improvements by using full covariance matrices. Now an average 6.22% and 17.72% improvement in R and RMSE, respectively, is achieved. If we look at the graph of the objective MOS *versus* subjective MOS (vide Fig. 2) we see that the estimates no longer are aligned with the axis, i.e. the correlation between the predictor and the response variables have been properly modelled.

IV. CONCLUSION

A novel objective speech quality measurement algorithm is proposed based on Gaussian mixture models. When using diagonal Gaussian components we have shown 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 technique were correlated and the use of only five diagonal Gaussian components was not enough to compensate for this. Still, if data is limited, diagonal components can be used with an average improvement in RMSE of 12.31%.

In the case a larger dataset is available, one is motivated to use full Gaussian components as we have shown an average improvement over PESQ of 6.22% and 17.72% in R and RMSE respectively.

REFERENCES

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

[2] 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.

[3] 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.

[4] W. Zha and W.-Y. Chan, "A data mining approach to objective speech quality measurement," to appear in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004.*

[5] 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.

[6] D. Ormoneit and V. Tresp, "Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging," in *Advances in Neural Information Processing Systems*, vol. 8. The MIT Press.