An Improved GMM-Based Voice Quality Predictor

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

A voice quality prediction method based on Gaussian mixture models (GMMs) is improved by constructing a feature selection algorithm to provide the best GMM-based prediction quality. The proposed sequential selection algorithm performs N-survivor search, allowing for trading between design complexity and performance. Simulation shows that predictors designed using the proposed algorithm outperform two benchmark selection algorithms. Performance improvements over the ITU-T P.862 PESQ standard are also attained.

1. Introduction

In [1], a novel method of speech quality estimation based on Gaussian mixture models (GMMs) is proposed. First, perceptual features are extracted from the distortion surface between an original speech signal and its degraded counterpart. Salient features are then selected using two statistical data mining methods, multivariate adaptive regression splines (MARS) [2] and classification and regression trees (CART) [3]. Lastly, features are mapped to a mean opinion score (MOS) [4] by means of a minimum mean squared error (MMSE) GMM-based estimator.

When designing GMM-based estimators, the features selected by CART or MARS may not lead to high estimation accuracy as the selection process is optimized for CART/MARS regressors. Indeed, in [1], diagonal covariance GMMs are shown to provide only modest performance and this is attributed to inherent characteristics of the features selected by CART or MARS. Here, we improve feature selection by proposing a feature selection algorithm whose selection criterion is the quality of the GMM-based estimator.

Simulation results show that the GMM-based estimators designed using the proposed algorithm better predict voice quality when compared to estimators trained on features selected by CART or MARS. Furthermore, an experiment performed on unseen data demonstrates that performance improvement over the International Telecommunications Union ITU-T P.862 standard, also known as Perceptual Evaluation of Speech Quality (PESQ) [5], is attained.

2. GMM-Based Voice Quality Prediction

2.1. Background

The goal in MMSE voice quality prediction is to find a set of features, represented by the feature vector $\mathbf{x}$, and a regression function $\hat{f}(\mathbf{x})$ that maps features to a predicted MOS. Both $\mathbf{x}$ and $\hat{f}(\mathbf{x})$ are chosen to minimize the mean squared error, ε_{MSE}, between $\hat{f}(\mathbf{x})$ and the subjective MOS (y), viz $\varepsilon_{MSE} = E[(y - \hat{f}(\mathbf{x}))^2]$. It is known that ε_{MSE} is minimized when $\hat{f}(\mathbf{x}) = E[y|\mathbf{x}]$, the conditional expectation of the subjective MOS, given $\mathbf{x}$. Before we introduce GMM-based estimators, a brief description of GMMs is given for the sake of notation.

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

$$p(\mathbf{u}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i . b_i(\mathbf{u}) \quad (1)$$

where $\alpha_i \geq 0, i = 1, ..., M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, $b_i(\mathbf{u})$, $i = 1, ..., M$ are the K-variate Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The Gaussian mixture density is parameterized by the elements $\boldsymbol{\lambda}_i = [\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i]$ which are estimated via the EM algorithm [6]. We use the *k-means* algorithm to initialize the GMM parameters.

The GMM-based estimators rely on modelling the joint density of $\mathbf{u} = [y, \mathbf{x}]^T$ with (1). Given the GMM parameters, the MMSE regression function is [7]

$$E[y|\mathbf{x}] = \sum_{i=1}^{M} h_i(\mathbf{x})[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(\mathbf{x} - \mu_j^x)]. \quad (2)$$

The above estimator is a weighted sum of linear models, where the weight $h_i(\mathbf{x})$ is the probability that the i^{th} Gaussian component generated the vector $\mathbf{x}$ and given by

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

Next, a description of the proposed feature selection algorithm is given.

2.2. Feature Selection

The proposed algorithm starts with an empty feature set and features from a candidate feature set are added to the set progressively. To determine which candidate feature to add, the algorithm tentatively adds to the current feature set one feature that is not already selected to form an augmented feature set. The joint density of the target variable and the augmented feature set is modelled with a GMM, with model parameters $\boldsymbol{\lambda}$ estimated using the EM algorithm. The accuracy of the GMM estimator using $\boldsymbol{\lambda}$ is then calculated. The above is repeated for every candidate feature and corresponding GMM. The candidate feature that produces the least regression error is admitted into the current feature set to form an updated feature set. The algorithm stops when the desired number of features has been selected. Note that the proposed algorithm progressively constructs $\hat{f}$ as features are being selected.

It is worth mentioning that for each candidate feature the best number of Gaussian components in (1) can be determined by checking different values of M. Using the notation "EM" to stand for GMM parameter estimation via the EM algorithm, $\hat{f}_k$ for the mapping function with k variables, and D for the desired number of features, the algorithm can be summarized as follows:

(0) Let $I = \{1, \ldots, n\},\ S = \emptyset,\ k = 1$;
(1) $\boldsymbol{\lambda}_i \leftarrow \text{EM}(y,\ S \cup \{x_i\}), \forall i \in I$;
(2) $i_k = \arg\min_{i \in I} \sum_j (y_j - \hat{f}_k(S \cup \{x_i\}|\boldsymbol{\lambda}_i))^2$;
(3) $I \leftarrow I - \{i_k\},\ S \leftarrow S \cup \{x_{i_k}\},\ k \leftarrow k + 1$.
(4) Go to step 1, stop if $k > D$.

2.3. N-Survivor Search

With a corresponding increase in computational complexity, the algorithm can perform sequential multiple-survivor search. So far, the algorithm description has focused on one survivor, i.e., the one feature variable that minimizes estimation error. In N-survivor search, at each iteration, the N features that assume the top-N ranks in minimizing the estimation error are kept as "survivors." A tradeoff between complexity and performance can be adjusted by tuning the parameter N.

If the ultimate goal is to find D features out of n candidate features, then N survivors are kept in iterations $i = 1, 2, \ldots, D-1$. At iteration $i = 1$, the algorithm selects the N best features out of the n available candidates. At iterations $1 < i < D$ the N best ranked features, out of the $N(n - i + 1)$ possible feature combinations, are kept. Lastly, at iteration $i = D$, the single best feature is kept. The last best feature and its ancestor features constitute the set of features selected by the search process.

The next section is dedicated to testing the accuracy of the proposed algorithm. Comparisons with GMM estimators trained on features selected by CART or MARS are carried out in the first experiment. Comparisons with PESQ are shown in the second, and an estimation test with unseen data is described in the third experiment.

3. Performance Results

The GMM for speech quality estimation is built on perceptual feature variables obtained by classifying perceptual distortions under a variety of contexts to form a pool of 209 candidate features [1]. Thirteen MOS labelled speech databases are used, containing a total of 5864 speech files. We use 10-fold cross validation to provide robustness in the performance evaluation. Estimation performance is assessed by the correlation (R) between subjective MOS and estimated MOS and by root-mean-square MOS error ($RMSE$).

3.1. Experiment I

The first experiment compares GMM estimators trained on features selected by our proposed feature selection algorithm to estimators trained on features selected by CART or MARS. For this experiment we check all permissible values of M at each iteration. To allow comparisons with [1] we search for $D = 5$ features. We restrict $M \leq 5$ in order to maintain an adequate training ratio (ratio between the number of parameters that have to be estimated and the total number of files in the training set) of 37 for full covariance matrices and 81 for diagonal matrices.

Let M_i be the number of Gaussian components chosen in iteration i of the proposed algorithm, it was found that the following combinations were often selected throughout the ten cross validation trials:

- Diagonal: $M_1 = 4,\ M_2 = M_3 = M_4 = M_5 = 5$;
- Full: $M_1 = 2,\ M_2 = 3,\ M_3 = M_4 = 4,\ M_5 = 5$.

Note that over the five algorithm iterations (D=5) used in this experiment the number of Gaussian components either increases or stays the same as the algorithm progresses. As expected, full covariance GMMs use fewer Gaussian components at the beginning, and the number of components increases with the number of features.

Figures 1 (a) and (b) compare performance figures for a 5-component GMM estimator designed using the proposed algorithm to that of an estimator designed using CART or MARS, for diagonal and full covariance matrices, respectively. Note that the proposed algorithm achieves higher R and lower $RMSE$ for all ten cross validation trials.

More precisely, if the percentage improvement in R is defined as

$$\% \uparrow R = \frac{R_{new} - R_{old}}{1 - R_{old}} \times 100\% \qquad (4)$$

where R_{new} and R_{old} are the correlation obtained using the proposed method and using CART or MARS, respec-

tively; diagonal GMM estimators incur an average improvement in R of 26.95% and 38.94 % when compared to CART and MARS, respectively. An average improvement of 31.10% and 20.01% is achieved for full GMM estimators. In turn, diagonal predictors trained on the proposed algorithm reduce $RMSE$ by an average of 13.93% and 24.16% when compared to CART and MARS, respectively. An average decrease of 19.07% and 11.96% is obtained for full covariance GMMs.

If multiple survivor search is carried out, performance can be improved. There is, however, a linear increase in design complexity. The 1-survivor algorithm needs to invoke the EM algorithm $M\sum_{i=1}^{D}(n-i+1)$ times, n being the total number of candidate features and D the desired number of features to be selected. Here, $n = 209$ and $D = 5$. By using the N-survivor approach, the number of EM invocations increases to $NM\sum_{i=1}^{D}(n-i+1)$. A simple experiment is carried out with $N = 2$ and simulations show that an improvement of 7.21% in R and a reduction of 3.12% in $RMSE$ can be attained by using 2-survivor search relative to single-survivor search.

3.2. Experiment II

In this experiment we compare performance of the GMM-based voice quality predictor to the performance of PESQ with the mapping proposed in [8]. Table 1 summarizes the performance figures; the column labelled "$\uparrow\%R$" shows improvement in R relative to PESQ by using a 5-component GMM estimator, trained with features selected by the proposed algorithm. Similarly, "$\downarrow\%RMSE$" denotes decrease in $RMSE$ relative to PESQ. Full GMM estimators outperform PESQ by 26.12% and 18.04% in R and $RMSE$, respectively. With 2-survivor search, an average improvement of approximately 29% in R and an average decrease of 19.51% in $RMSE$ is attained. Additionally, it is important to note that, despite lower performance, full GMM estimators trained on features selected by CART or MARS also outperform PESQ, as was shown in [1].

3.3. Experiment III

In this last experiment, the proposed algorithm is tested on unseen data, i.e., data that has not been used in the training of the GMM predictors. Two unseen test databases are used, each comprised of approximately 3000 subjectively scored speech file pairs, with speech under various degradation conditions. For both databases, the proposed algorithm achieves an average 5% lower correlation when compared to PESQ. However, for the first database, the proposed algorithm reduces $RMSE$ by an average 41%. For the second database, an average decrease of 19% is attained. It is important to realize that $RMSE$ is a more realistic measure of estimator performance. It can be shown that $RMSE$ is the sum of unexplained variance in the regression model, MOS estimation error due to limited number of listeners (affecting all algorithms equally), and bias error between subjective MOS and objective MOS. The calculation of R does not take into consideration this bias error [1].

4. Conclusion

We have proposed a feature selection algorithm for speech quality assessment based on Gaussian mixture models. The algorithm provides for trading between complexity and performance by adjusting the number of survivors searched. Simulation results show that GMM estimators designed using the proposed algorithm outperform two benchmark selection algorithms, with N-survivor search providing better performance. Furthermore, a test on unseen data shows that the proposed algorithm reduces $RMSE$ by an average 32% relative to PESQ.

5. References

[1] T. H. Falk, W.-Y. Chan, and P. Kabal, "Speech quality estimation using Gaussian mixture models," in *Proc. of the Int. Conf. on Spoken Language Processing*, Oct. 2004, pp. 2013–2016.

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

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

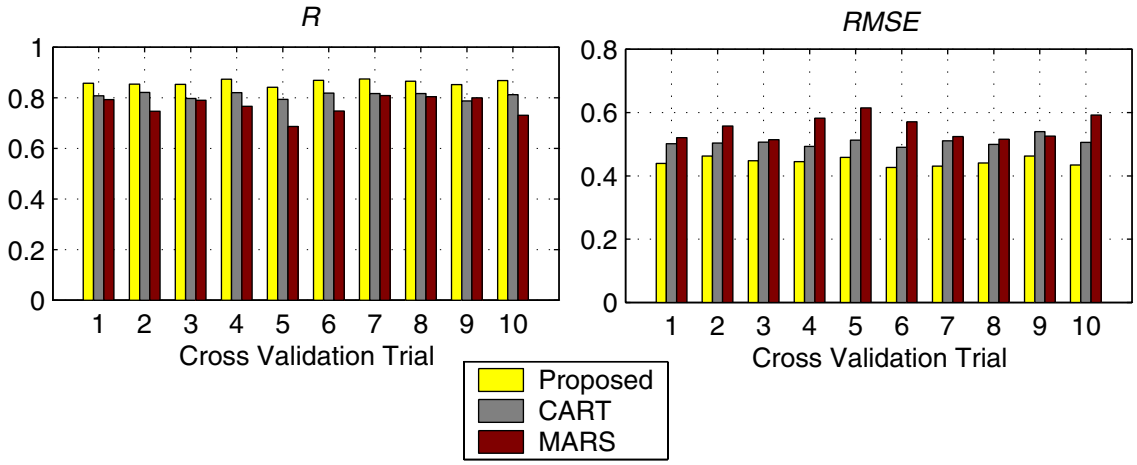
[4] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," International Telecommunication Union, Geneva, Switzerland, Feb. 1996.

[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, Feb. 2001.

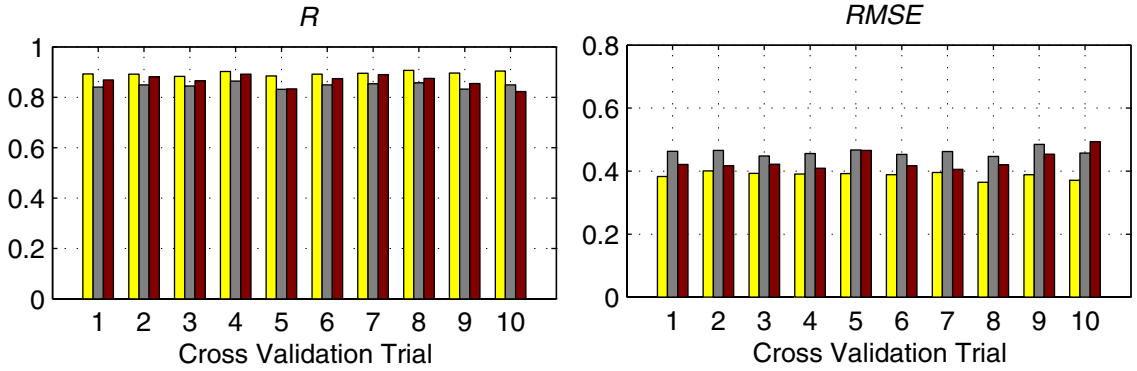
[6] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[7] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, vol. 6, 1994, pp. 120–127.

[8] ITU-T P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," International Telecommunication Union, Geneva, Switzerland, Nov. 2003.



(a)



(b)

Figure 1: Correlation R and $RMSE$ comparisons between GMM estimators trained on features selected by CART, MARS, and the proposed algorithm for (a) diagonal and (b) full GMM estimators.

Table 1: Performance comparison: PESQ and proposed algorithm

Cross Validation	PESQ		1-Survivor (diagonal)		1-Survivor (full)		2-Survivor (full)	
Trials	R	$RMSE$	$\uparrow\%R$	$\downarrow\%RMSE$	$\uparrow\%R$	$\downarrow\%RMSE$	$\uparrow\%R$	$\downarrow\%RMSE$
Trial 1	0.8568	0.4643	0.70	5.44	25.35	17.51	29.26	19.53
Trial 2	0.8535	0.4871	0.27	5.09	26.08	17.78	26.08	17.78
Trial 3	0.8460	0.4809	4.55	6.86	24.35	18.27	29.81	20.84
Trial 4	0.8670	0.4670	4.66	4.75	26.54	16.33	30.23	18.24
Trial 5	0.8449	0.4811	-2.13	4.69	25.98	18.46	28.18	19.58
Trial 6	0.8564	0.4668	9.05	8.61	24.72	16.69	28.90	18.81
Trial 7	0.8738	0.4633	0.16	7.08	17.04	14.63	22.35	17.09
Trial 8	0.8581	0.4801	5.29	8.16	34.81	24.09	34.81	24.09
Trial 9	0.8608	0.4695	-6.25	1.53	25.50	17.17	27.51	18.23
Trial 10	0.8623	0.4604	3.92	5.71	30.79	19.46	33.55	20.98
Average			**2.02**	**5.79**	**26.12**	**18.04**	**29.07**	**19.52**