BAYESIAN MODEL BASED NON-INTRUSIVE SPEECH QUALITY EVALUATION

Guo Chen, Vijay Parsa

National Centre for Audiology, Elborn College,
Dept. of Electrical & Computer Engineering, University of Western Ontario, N6G 1H1, Canada
E-mail: sguo@nca.uwo.ca, parsa@nca.uwo.ca

ABSTRACT

A novel Bayesian model-based non-intrusive speech quality evaluation (BM-NiSQE) algorithm is presented in this paper. The proposed BM-NiSQE algorithm employs a statistical model approach and Bayesian inference to estimate the speech quality only using the output signal of the system under test. In the proposed algorithm, the speech features are extracted by the perceptual spectral analysis. Gaussian mixture density hidden Markov models (GMD-HMMs) are exploited to characterize different speech quality categories, which take into account not only the temporal variations of speech signal but also the spectral statistical characteristics in the perception domain. Based on the trained GMD-HMMs, the prediction of speech quality is carried out by the Bayesian inference and minimum mean square error (MMSE) estimation. Preliminary experimental results show that the predicted results of the proposed algorithm correlate well with the subjective quality scores.

1. INTRODUCTION

One of the most important criteria for the performance of compression and transmission technologies in speech communication systems is subjective speech quality, i.e., the users' perception of the quality of the received signals. The widely used subjective speech quality testing method is the opinion rating scheme as defined in the ITU-T Recommendation P.800 [1]. The subjective speech quality is evaluated by an absolute category rating (ACR) scale in which the quality is represented by five grades - excellent(5), good(4), fair(3), poor(2), and bad(1). Typically, the ratings are collected from listeners and the arithmetic mean of these scores is calculated to form the mean opinion score (MOS). While subjective opinions of speech quality are preferred as the most trustworthy criterion for speech quality evaluation, they are also time-consuming and expensive. Therefore, objective quality evaluation methods for predicting subjective scores of speech quality from physical parameters are desirable.

In the past two decades, objective quality evaluation has received considerable attention and is still an active research topic [2, 3, 4, 5, 6]. A majority of these objective methods are based on input/output comparisons, i.e., intrusive methods, which estimate the speech quality by measuring the "distortion" between the input and output signals, and mapping the distortion values to the predicted quality metric. But in some applications, a reference signal might not be available for an input/output comparison, e.g., wireless communications, voice over IP network, and hearing aids. In such cases, an attractive alternative approach is to assess speech quality using only the output signal, i.e., a non-intrusive speech quality evaluation [7]. An effective non-intrusive quality measure will be of significant importance to applications where an input signal is not readily available, such as the performance evaluation of hearing aids [8] and non-intrusive performance monitoring of communication systems[9]. Non-intrusive evaluation techniques recently described in the literature [10, 11], however, are mainly based on comparing the signal under test to an artificial reference signal representing the closest match from an appropriately formulated codebook. In this paper, we propose a novel non-intrusive evaluation algorithm based on a statistical model approach and Bayesian inference which does not require an artificial reference or codebook and operates on just the signal under test. The proposed algorithm uses pattern classification methodology to estimate speech quality, which is a completely different strategy in dealing with speech quality evaluation from the existing methods, such as the MPSDD algorithm [5]. In addition, the approach presented in this paper provides discrete speech quality ratings (in line with the subjective data) rather than a continuous speech quality variable.

It is well-known that speech signal not only provides the words or message being pronounced, but also gives information of sound quality perceived by the listeners as stated in a subjective listening test [1]. While speech recognition focuses on understanding the underlying text and information in the speech signal, speech quality evaluation is concerned with extracting the quality information from the sound signals. Success in correctly predicting subjective speech quality depends on extracting and modelling the characteristics of the speech signal that can effectively distinguish between different speech quality categories, such as the categories of the MOS scale. The proposed BM-NiSQE algorithm exploits perceptual spectral analysis to extract the speech features, uses GMD-HMMs to model the characteristics of different quality categories, and applies Bayesian inference and MMSE estimation to predict the subjective quality scores. The use of GMD-HMMs is motivated by the fact that the GMD-HMMs have a greater ability to capture both the temporal and spectral variabilities of speech signal in the perception domain.

2. OUTLINE OF THE BM-NISQE ALGORITHM

The block diagram of the proposed BM-NiSQE algorithm is shown in Fig.1. We assume that the subjective quality scores are classified into Q categories in our study. The subjective quality score associated with each quality category is denoted by $Y^{(q)}$, $q = 1, 2, ..., Q$. The observation features representing the q-th quality category, denoted by $O^{(q)}$, are extracted by the perceptual spectral analysis (a detailed description is given in Section III). The observation fea-

This work has been financially supported by the Oticon Foundation, Denmark and the Ontario Rehabilitation Technology Consortium, Canada.

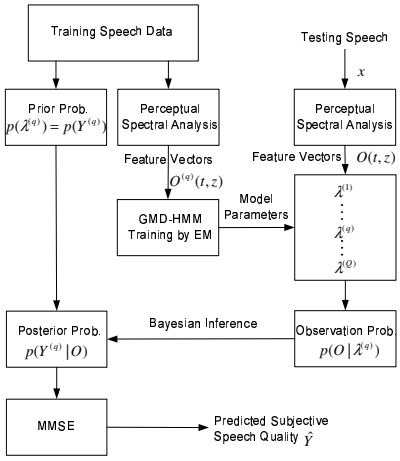

Fig. 1. The block diagram of the BM-NiSQE algorithm

tures $O^{(q)}$ are characterized by a GMD-HMM $\lambda^{(q)}, q = 1...Q$. The predicted subjective quality scores $\hat{Y}$ can be estimated using the minimum mean square error (MMSE) criterion.

2.1. MMSE estimation of subjective quality scores

The aim of the BM-NiSQE algorithm is to minimize the mean squared error between the predicted MOS values $\hat{Y}$ and the true MOS values Y, i.e.,

$$E[(\hat{Y} - Y)^2|O] \quad \rightarrow \quad \min, \tag{1}$$

where O is the observation features of the speech under test. Obviously, the solution of Eq.(1) is a conditional expectation as

$$\hat{Y} = E[Y|O] = \int Y p(Y|O) dY. \tag{2}$$

Since there are Q quality categories, Eq.(2) becomes

$$\hat{Y} = E[Y^{(q)}|O] = \sum_{q=1}^{Q} Y^{(q)} p(Y^{(q)}|O). \tag{3}$$

2.2. Bayesian Inference

Assuming that the speech signal of the i-th quality category is characterized by the GMD-HMM, $\lambda^{(i)}$, we have

$$p(Y^{(q)}|O) = \sum_{i=1}^{Q} p(Y^{(q)}|\lambda^{(i)}) p(\lambda^{(i)}|O). \tag{4}$$

Using Bayesian inference, the posterior probability can be reformulated by the prior probability, $p(\lambda^{(i)})$, and the likelihood, $p(O|\lambda^{(i)})$, as below,

$$p(Y^{(q)}|O) = \sum_{i=1}^{Q} p(Y^{(q)}|\lambda^{(i)}) \frac{p(O|\lambda^{(i)}) p(\lambda^{(i)})}{\sum_{m=1}^{Q} p(O|\lambda^{(m)}) p(\lambda^{(m)})}. \tag{5}$$

Since the speech signal of the q-th quality category is associated with the corresponding statistical model, we have

$$p(Y^{(q)}|\lambda^{(i)}) = \begin{cases} 1, & i = q \\ 0, & i \neq q \end{cases} \tag{6}$$

Consequently, the following formula for the posterior probability is obtained

$$p(Y^{(q)}|O) = \frac{p(O|\lambda^{(q)}) p(\lambda^{(q)})}{\sum_{m=1}^{Q} p(O|\lambda^{(m)}) p(\lambda^{(m)})} \tag{7}$$

2.3. Estimation of the likelihood and prior probability

Obviously, the prior probability $p(\lambda^{(q)})$ can be estimated from the training data set, i.e., the probability of each existing quality category. As for the likelihood $p(O|\lambda^{(q)})$, due to the high dimensionality of the observation sequence O, it is not practical to estimate the joint conditional probability directly from the speech signal, e.g. using a histogram method. However, using a parametric model, such as a GMD-HMM, makes estimation from data feasible. The problem of estimating $p(O|\lambda^{(q)})$ is replaced by the problem of estimating the GMD-HMM parameters. The description of the GMD-HMMs employed by the BM-NiSQE algorithm is given in the Setion IV. After obtaining the likelihood and prior probability, the posterior probability $p(Y^{(q)}|O)$ can be calculated. Subsequently, the predicted subjective quality score can be estimated in terms of Eq.(3).

3. FEATURE EXTRACTION BY PERCEPTUAL SPECTRAL ANALYSIS

Previous research has shown that the time-frequency distribution of the speech signal is useful for quality evaluation[2], although there are no specific features to distinguish different speech quality categories. This is because the time-frequency distribution reflects different vocal tract configurations which produce the speech signal associated corresponding speech quality categories. Recent studies have found speech features combined with perceptual models of human's auditory system to be more effective in speech quality evaluation [4, 5, 6]. In the BM-NiSQE algorithm, we employ perceptual spectral features derived from speech time-frequency distribution to represent the quality characteristics of speech signal. The perceptual spectral analysis is performed by time-frequency mapping and transformation to the perception domain. The block diagram of the perceptual spectral analysis is shown in Fig.2. Typically, the input speech signal with an 8 kHz sampling frequency is first segmented into frames of 32 ms with an overlap of 50%. Each frame is then transformed to the frequency domain using a Hanning window and a short term FFT. The transformation to perception domain, i.e., Bark domain, is done by grouping the spectral coefficients into critical Bark bands. The perceptual spectral analysis is formulated in the following steps.

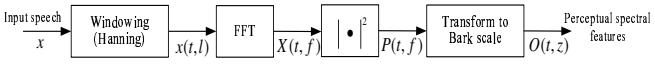

Fig. 2. The block diagram of the perceptual spectral analysis

3.1. Time-frequency mapping

The mapping from time domain to time-frequency domain is implemented by a short-term Fourier transform with a Hanning window resulting in a time-frequency representation with a constant resolution in both time and frequency domain. Let the t-th frame of speech signal be denoted by samples $x(t,l)$ with l running from 0 to $L-1$, where $t = 1, 2, ..., T$. The signal $x(t,l)$ is windowed by a Hanning window, i.e.,

$$x_w(t,l) = h(l)x(t,l), \tag{8}$$

where $h(l)$ is the windowing function as

$$h(l) = 0.5\left(1 - cos\left(\frac{2\pi l}{L}\right)\right), \quad 0 \le l \le L-1. \tag{9}$$

The Fast Fourier transform (FFT) is performed on $x_w(t,l)$ with $L = 256$ samples. The real and imaginary components are squared and added to get the short-term spectral power density $P(t,f)$, i.e.,

$$P(t,f) = (\text{Re } X_w(t,f))^2 + (\text{Im } X_w(t,f))^2, \tag{10}$$

where f is frequency scale.

3.2. Transform to perception domain

The spectral power density $P(t,f)$ is warped from the Hertz scale to the critical band scale (i.e., Bark scale), leading to a perceptual spectral density representation within each frame. $P(t,f)$ is partitioned into critical bands and the energies of each critical band are added up. The conversion from frequency to Bark scale is

$$z = 7 \cdot asinh(f/650). \tag{11}$$

In order to improve the resolution of the perceptual spectral density distribution, we chose an interval of $\frac{1}{2}$ Bark scale. This leads to 29 critical bands covering 300 Hz - 4 kHz frequency. The energy in each critical band is summed as follows.

$$O(t,z) = \sum_{v=bl_z}^{bh_z} P(t,v), \tag{12}$$

where bl_z and bh_z are the lower and upper boundaries of critical band z, respectively. The grouped energies are denoted by $O(t,z), z = 1, 2, ..., 29$. Each speech frame is therefore represented by a vector of 29 perceptual spectral features.

4. POSTERIOR PROBABILITY ESTIMATION BY GMD-HMMS

In the BM-NiSQE algorithm, the joint conditional observation densities $p(O|\lambda^{(q)})$ $(q = 1, 2, ..., Q)$ are estimated by GMD-HMMs. There are Q such GMD-HMMs $\lambda^{(q)}$, where $\lambda^{(q)} = \{\pi^{(q)}, A^{(q)}, B^{(q)}\}$ denotes the set parameters of the q-th N-state GMD-HMM used to characterize the q-th speech quality category, of which $\pi^{(q)}$ represents the initial state distribution, $A^{(q)}$ is the transition probability matrix, and $B^{(q)}$ is the parameter vector composed of mixture parameters $B_i^{(q)} = \{\omega_{ik}^{(q)}, \mu_{ik}^{(q)}, \sigma_{ik}^{(q)}\}$ for state i. The state observation probability density function (pdf) at state i is assumed to be a mixture of multivariate Gaussian pdfs:

$$p(B_i^{(q)}) = \sum_{k=1}^{K} \omega_{ik}^{(q)} \mathcal{N}(O; \mu_{ik}^{(q)}, \Sigma_{ik}^{(q)}) \tag{13}$$

where O is the observed features, the set of mixture coefficients $\omega_{ik}^{(q)}$ satisfy the constraint $\sum_{k=1}^{K} \omega_{ik}^{(q)} = 1$, and $\mathcal{N}(O; \mu_{ik}^{(q)}, \Sigma_{ik}^{(q)})$ is the k-th Gaussian mixture component with $\mu_{ik}^{(q)}$ being the D-dimensional mean vector and $\Sigma_{ik}^{(q)}$ being the $D \times D$ covariance matrix, i.e.

$$\mathcal{N}(O;\mu_{ik}^{(q)}, \Sigma_{ik}^{(q)}) = \frac{1}{(2\pi)^{D/2}|\Sigma_{ik}^{(q)}|^{1/2}} \times \exp\left(-\frac{1}{2}(O - \mu_{ik}^{(q)})'(\Sigma_{ik}^{(q)})^{-1}(O - \mu_{ik}^{(q)})\right), \tag{14}$$

where the apostrophe represents a transpose operation. For each state i, the complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Also, it is assumed that all the state observation pdfs have the same number of mixture components. Correspondingly, the q-th GMD-HMM can be represented the parameters collectively by the notation

$$\lambda^{(q)} = \{\pi^{(q)}, A^{(q)}, \omega^{(q)}, \mu^{(q)}, \Sigma^{(q)}\}, q = 1, ..., Q. \tag{15}$$

Each speech quality category is represented by a GMD-HMM and is referred to by the model $\lambda^{(q)}$.

The choice of GMD-HMMs is based on the attributes that these models have been proven extremely useful in modelling pdfs of speech signal [12]. The GMD-HMMs are capable of capturing the time-frequency variabilities of the speech signal and representing the properties of general vocal tract configurations that are quite useful for characterizing speech quality [7]. Furthermore, a linear combination of GMD functions has a great ability to form smooth approximations to arbitrarily-shaped densities.

The training of the parameters of the GMD-HMMs, i.e., π,A, ω,μ and Σ, can be performed with the *expectation-maximization* (EM) algorithm. A starting point for the iterative refinement of the EM algorithm is determined by clustering the training data with the K-means algorithm in our study. One critical factor in training a GMD-HMM is the selection of the order K of the Gaussian mixture components and the order N of the hidden states. There are no good theoretical means to guide one in either of these selections, so they are best determined experimentally for a given task. An experimental examination of these factors is discussed in the following Section V.

5. PERFORMANCE EVALUATION

A speech data set with different MOS subjective quality scores was used to test the performance of the BM-NiSQE algorithm. This speech data set consists of 128 MOS values, including five modulated noise reference unit (MNRU) conditions (5-25 dB at 5 dB step) and different types of speech coders, such as ADPCM, GSM, IS54, FS1016, LD-CELP and CELP. The correlation coefficient (denoted by ρ) and standard error of estimate (denoted by ϵ), defined in [2], were used to evaluate the performance. Three-quarters of the data set was used as a training data set while the remaining was used as a testing data set. Based on the original categories of the MOS scale, we classified the experimental data set into nine categories denoted by ξ_i, where $\{1 \le \xi_1 < 1.25 \le \xi_2 < 1.75 \le \xi_3 < 2.25 \le \xi_4 < 2.75 \le \xi_5 < 3.25 \le \xi_6 < 3.75 \le \xi_7 < 4.25 \le \xi_8 < 4.75 \le \xi_9 \le 5\}$.

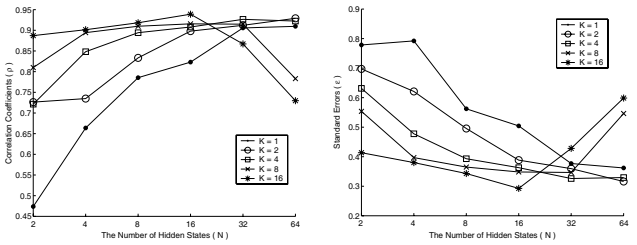

Fig. 3. Investigation of the selection of the model orders in terms of correlation coefficients and standard errors

5.1. Numerical computational issues

The BM-NiSQE algorithm was programmed in Matlab. In our numerical computation, the smallest positive floating point number is approximately 2.225×10^{-308}. With this precision, Equation (7) may sometimes cause numerical problems due to the term $p(O|\lambda^{(q)})$ which is significantly small and generally exceeds the precision range, resulting in an underflow problem. A remedy for this was to use a logarithmic value to calculate the joint condition probability $p(O|\lambda^{(q)})$ first and a scaling factor ϕ is then employed to calculate the final posterior probability $p(Y^{(q)}|O)$ as follows:

$$p(Y^{(q)}|O) = \frac{10^{\left(\log p(O|\lambda^{(q)}) - \phi\right)} p(\lambda^{(q)})}{\sum_{m=1}^{Q} 10^{\left(\log p(O|\lambda^{(m)}) - \phi\right)} p(\lambda^{(m)})} \qquad (16)$$

where

$$\phi = \max_{m} \log p(O|\lambda^{(m)}), \quad m = 1, 2, ..., Q$$

5.2. Selection of the model orders

As indicated in Section IV, the selection of the orders of hidden states and Gaussian mixture components is critical in the BM-NiSQE algorithm. Choosing too low an order will produce a model which does not accurately capture the distinguishing characteristics of different quality categories. Choosing too high orders will reduce performance when there are a large number of model parameters relative to the available training data and will also result in excessive computational complexity both in training and testing. To investigate the performance of the BM-NiSQE algorithm with respect to the model orders, the following experiments were conducted on the training data set. The GMD-HMMs with 2, 4, 8, 16, 32, 64 hidden states and 1, 2, 4, 8, 16 Gaussian components were trained using the data set. Fig.3. shows the results of the performance versus the selected orders in terms of the correlation coefficients and standard errors. From the results, it can be seen that the GMD-HMMs with 16 hidden states and 16 Gaussian components attained the best performance, and that the correlation coefficients reached 0.9386 with a standard error of 0.2922.

5.3. Results of the testing data

The experiments on the testing data were carried out by the GMD-HMMs with $K = 16$ and $N = 16$. The correlation coefficients between the predicted scores and the true MOS values attained 0.8962, and the standard error of prediction was 0.3974. These results demonstrated the ability of the proposed BM-NiSQE algorithm to predict speech quality scores without access to the input speech signal. This compares favorably with the standard intrusive speech quality measure, PESQ [3], which provides a correlation of 0.9265.

6. CONCLUSIONS

In this paper a novel Bayesian model based non-intrusive speech quality evaluation algorithm is proposed. The proposed BM-NiSQE algorithm is based on the perceptual spectral features, and applies Bayesian inference and Gaussian mixture density hidden Markov models to predict the subjective quality scores without any reference signal. Preliminary experimental results show that the predicted results of the proposed algorithm correlate well with the subjective quality scores.

7. REFERENCES

[1] ITU-T, "Methods for subjective determination of transmission quality," *ITU-T P.800 Recommendation*, Aug. 1996.

[2] S.R.Quackenbush, T.P.Barnwell-III, and M.A.Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ., 1988.

[3] ITU-T, "Perceptual evaluation of speech quality," *ITU-T P.862 Recommendation*, Feb. 2001.

[4] G.Chen, S.Koh, and I.Soon, "Enhanced itakura measure incorporating masking properties of human auditory system," *Signal Processing*, vol. 83, pp. 1445–1456, 2003.

[5] G.Chen and V.Parsa, "Output-based speech quality evaluation by measuring perceptual spectral density distribution," *IEE Electronics Letters*, vol. 40, no. 12, pp. 783–784, Jun. 2004.

[6] A.Rix, "Perceptual speech quality assessment - a review," *Proceedings of ICASSP-2004*, vol. 3, pp. 1056–1059, May 2004.

[7] P.Gray, M.P.Hollier, and R.E.Massara, "Non-intrusive speech quality assessment using vocal-tract models," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 147, no. 6, pp. 493–501, Dec 2000.

[8] L.B.Nielsen, *Objective scaling of sound quality for normal-hearing and hearing-impaired listerners*, Technical report, Report No.54, The acoustics laboratory, Technical University of Denmark, 1993.

[9] D.S.Kim and A.Tarraf, "Perceptual model for non-intrusive speech quality assessment," *Proceedings of ICASSP-2004*, vol. 3, pp. 1060–1063, May 2004.

[10] D.Picovici and A.E.Mahdi, "Output-based objective speech quality measure using self-organizing map," *IEEE Proceedings of ICASSP-2003*, vol. 1, pp. 476–479, 2003.

[11] D.Picovici and A.E.Mahdi, "New output-based perceptual measure for predicting subjective quality of speech," *Proceedings of ICASSP-2004*, vol. 5, pp. 633–636, May 2004.

[12] G.Chen and V.Parsa, "Hmm-based frequency bandwidth extension for speech enhancement using line spectral frequencies," *Proceedings of ICASSP-2004*, vol. 1, pp. 709–712, May 2004.