OUTPUT-BASED SOUND QUALITY EVALUATION USING STATISTICAL MODEL METHOD

Guo Chen, Vijay Parsa
National Centre for Aduiolgy, Faculties of Health Sciences & Engineering, Elborn College,
University of Western Ontario, Ontario, Canada, N6G 1H1. Email: sguo@nca.uwo.ca, parasa@nca.uwo.ca

1. INTRODUCTION

Good objective speech quality measurement is highly desirable and useful in the field of speech communication. The accuracy of the objective measurement is determined by correlating it with known subjective measurement of speech quality, such as the mean opinion score (MOS) defined in [1]. A majority of previously reported objective methods are based on input/output comparisons, 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., evaluation of pathological voice. In such cases, an attractive alternative is to assess speech quality using only the output signal, i.e., output-based evaluation. Currently, a handful of output-based evaluation techniques have been reported in the literature. In [2], a vocal tract modeling based non-intrusive evaluation technique was presented to monitor telecommunication network distortion. In [3], an auditory non-intrusive quality estimation (ANIQUE) model, which is based on human auditory and articulation systems, was reported. In [4-6], the authors investigated output-based techniques by measuring perceptual spectral density distribution and by exploiting neuro-fuzzy techniques. In [7], a state-of-the-art non-intrusive evaluation standard was recommended by the ITU, which applies a complicated evaluation structure and myriad parameters. Clearly, it can be seen that all of the current output-based techniques reported in the literature are intended to provide an optimal continuous mapping from physical parameters to subjective speech quality scores (such as the MOS scores). But this is inconsistent with actual subjective listening-opinion tests [1]. As stated in [1], the subjective MOS listening-opinion test is a Category Judgment method (Page 3 in [1]), which can be thought of as a process of speech quality pattern classification. Therefore, it is intuitive to develop an evaluation method by employing statistical pattern classification to imitate subjects' behavior in a subjective listening-opinion test. In this paper, we propose a novel output-based evaluation technique using statistical modeling, which is in line with a subjective listening-opinion test. The proposed method was motivated by two facts: (i) the subjects in a listening test do in fact perform a quality pattern classification process in terms of their perception of the speech signal, and (ii) in the process of a listening test, the subjects listen only to the output speech signal of the system under test, not hear the original input signal (i.e., the reference signal).

2. THE PROPOSED METHOD

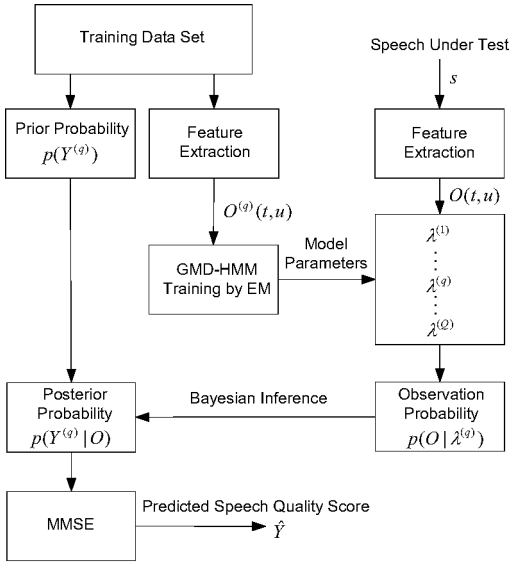

Fig.1. The Block Diagram of The Proposed Method

The block diagram of the proposed method is shown in Fig.1. We assume that subjective quality scores are classified into Q categories, 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 pitch power density analysis. The observation features $O^{(q)}$ are characterized by GMD-HMMs $\lambda^{(q)}, q=1,...,Q$. The aim of the proposed method is to minimize the mean squared error between the predicted MOS values $\hat{Y}$ and the true MOS values Y given observations, i.e., let $E[(\hat{Y}-Y)^2|O]$ approaches to minimum. The solution is a conditional expectation as $\hat{Y}=E[Y|O]=\int Yp(Y|O)dY$. Since there are Q quality categories, The conditional expectation becomes $\hat{Y}=E[Y|O]=\sum_{q=1}^{Q}Y^{(q)}p(Y^{(q)}|O)$. The conditional probability $p(Y^{(q)}|O)$ can be calculated by

$$p(Y^{(q)} \mid O) = \sum_{i-1}^{Q} p(Y^{(q)} \mid \lambda^{(i)}) p(\lambda^{(i)} \mid O)$$

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

$$p(Y^{(q)} \mid O) = \sum_{i=1}^{Q} p(Y^{(q)} \mid \lambda^{(i)}) \frac{p(O \mid \lambda^{(i)}) p(\lambda^{(i)})}{\sum_{m=1}^{Q} p(O \mid \lambda^{(m)}) p(\lambda^{(m)})} \quad (1)$$

Since we assumed that the speech signal of the *q-th* quality category is generated by the corresponding statistical model *i*, we have $p(Y^{(q)} \mid \lambda^{(i)}) = 1$ if $i = q$ and $p(Y^{(q)} \mid \lambda^{(i)}) = 0$ if $i \neq q$. Therefore, we may rewrite Eq.(1) as

$$p(Y^{(q)} \mid O) = \frac{p(O \mid \lambda^{(q)}) p(\lambda^{(q)})}{\sum_{m=1}^{Q} p(O \mid \lambda^{(m)}) p(\lambda^{(m)})} \quad (2)$$

In Eq.(2), the prior probability can be estimated from the training data, i.e., the probability of each existing quality category. The likelihood is estimated by a GMD-HMM.

In the proposed method, we employ the spread pitch power density values as the observation features for representing different speech quality categories. The feature extraction process consists of the following steps: (1) Voice Activity Detection; (2) Level normalization & IRS filtering; (3) Time-frequency mapping; (4) Outer and middle ear transfer function; (5) Transform to pitch (Bark) domain; (6) Adding internal noise; (7) Frequency domain spreading; Finally, the logarithmic values of the spread pitch power density are chosen as the observation features for each speech frame, denoted by *O(t,u)*.

The joint conditional observation densities $p(O \mid \lambda^{(q)})$ 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 training of the parameters of the GMD-HMMs is performed with the *expectation-maximization* (EM) algorithm. A starting point is determined by clustering the training data with the *K*-means algorithm in our study.

3. EXPERIMENTAL RESULTS

The experimental data consist of seven subjective quality MOS databases obtained in two listening opinion tests as described in Experiment One and Three of the ITU-T P-Series Supplement 23[8]. We combined these seven databases into a global database, giving a total of 1328 MOS scores. Three-quarters of the global database was used for training, while the remaining was used as a validation data set. The correlation coefficient (ρ) and standard error of estimate (ε) were used to evaluate the performance. The results are shown in Fig.2. From the results, it can be observed that the correlation of the proposed method attained 0.9012 with a standard error of 0.3390 across the whole databases. This compares favorably with the P.563, which provides a correlation and standard error of 0.8422 and 0.4493, respectively. The results demonstrated the ability of the proposed method to predict speech quality scores without any reference signal.

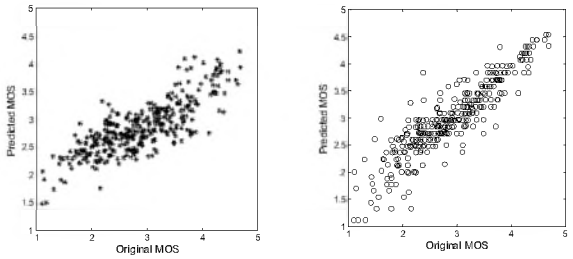

Fig.2. Predicted MOSs. *: P.563; o: Proposed Method

4. REFERENCES

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

[2] 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, 2000.

[3] D.-S. Kim, "Anique: An auditory model for single-ended speech quality estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, 2005.

[4] 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, 2004.

[5] G. Chen and V. Parsa, "Non-intrusive speech quality evaluation using an adaptive neuro-fuzzy inference system," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 403–406, 2005.

[6] G. Chen and V. Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *Proc. IEEE of ICASSP* 2005, pp. 385–388, 2005.

[7] *Single ended method for objective speech quality assessment in narrow-band telephony applicaitons*, ITU-T Recommendation P.563, 2004.

[8] *ITU-T coded-speech database*, ITU-T P-series Supplement 23, 1998.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge the financial support by the Oticon Foundation, Denmark, the Ontario Rehabilitation Technology Consortium, Canada, and the NSERC, Canada.