Evaluation of speech quality using digital watermarking

Libin Cai and Jiying Zhao[a)]

MCRLab, School of Information Technology and Engineering, University of Ottawa, 800 King Edward Ave., Ottawa, Ontario, Canada, K1N 6N5

a){*lcai,jyzhao*}*@site.uottawa.ca*

Abstract: Speech quality evaluation is a very important research topic. Mean Opinion Score (MOS) is reliable but the listening test is very expensive, time consuming, and sometimes impractical. The existing objective quality assessment methods require either the original speech or complicated computation model, which makes some applications of quality evaluation impossible. We propose to use digital audio watermarking to evaluate the quality of speech. Our method does not need original signal or computation model. The experimental results show that the method yields accurate quality scores which are very close to the results of PESQ.

Keywords: Watermarking, speech quality evaluation, MOS, PESQ

Classification: Science and engineering for electronics

References

[1] T. H. Falk and W.-Y. Chan, "Objective Speech Quality Assessment Using Gaussian Mixture Models," *Proc. 32nd Biennial Symposium on Communications*, pp. 169–171, June 2004.

[2] L. Ding and R. Goubran, "Assessment of effects of packet loss on speech quality in VoIP," *Proc. IEEE HAVE2003*, pp. 49–54, Sept. 2003.

[3] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE ICASSP*, pp. 749–752, May 2001.

[4] L. Cai and J. Zhao, "Audio Quality Measurement by Using Digital Watermarking," *Proc. IEEE CCECE2004*, pp. 1159–1162, May 2004.

1 Introduction

The evaluation of audio and speech quality is of critical importance in today's computer network control, e-commerce, and telephone networks, mainly because quality is a key determinant of customer satisfaction and key indication of computer network condition. Traditionally, the only way to measure the perception of quality of a speech signal was through the use of subjective testing [1], in which 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

DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004

is also very expensive and time consuming. Due to these reasons, objective methods have been developed in recent years, classified into two categories: signal based methods and parameters based methods [2]. Signal based methods use the reference and degraded signals as the input to the measurement, such as the state-of-the-art objective measurement algorithm, PESQ (perceptual evaluation of speech quality) [3]. Meanwhile, parameters based methods predict the speech quality through a computational model instead of using real measurement. For example, Falk and Chan [1] proposed an approach to objective speech quality measurements using Gaussian Mixture Models (GMMs). This kind of methods need a large training database to construct good estimators of subjective listening quality, and different training database may result in different model.

On the other hand, digital watermarking technology has been around for more than ten years, which has been used in copyright protection, content authentication, copy control, broadcast monitoring, etc. In this paper, we propose a new application of digital watermarking, speech quality evaluation. The basis of the method is that the carefully embedded watermark in a speech will suffer the same distortions as the speech does. The proposed method needs neither original speech, nor training database. Using PESQ as reference, the experimental results show that the proposed method gives very accurate quality evaluation. Furthermore, without the complicated signal processing on both original and degraded speeches, such as time alignment, equalization and FFT filtering, the implementation of the proposed quality evaluation is very fast. In addition to speech quality evaluation, this objective method can also evaluate the quality of audio.

2 Proposed speech quality evaluation method

The proposed method is based on Digital Wavelet Transform (DWT) and quantization. We embed watermark in the DWT coefficients of the speech. The watermark will undergo the same distortions as the speech does. Therefore, we can evaluate the quality of the speech having undergone distortions by evaluating the Percentage of Correctly Extracted Watermark bits (PCEW). The following introduces the watermark embedding and extraction, quality evaluation, and quantization step optimization, as illustrated in Fig. 1.

2.1 Watermark embedding and extraction

For watermark embedding, we compute the discrete wavelet transform of the speech signal, then we quantize the resulting coefficients with Equ. (1) [4]. By quantization, every real number is assigned a binary number 0 or 1.

$$Q(e) = \begin{cases} 0 & \text{if } k \times \Delta \le e < (k+1) \times \Delta \quad (k = 0, \pm 2, \pm 4, \ldots) \\ 1 & \text{if } k \times \Delta \le e < (k+1) \times \Delta \quad (k = 1, \pm 3, \pm 5, \ldots) \end{cases} \tag{1}$$

where e is the value of the coefficient, while Δ is a positive real number called quantization step. During the watermark embedding, after quantizing

© IEICE 2004
DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004

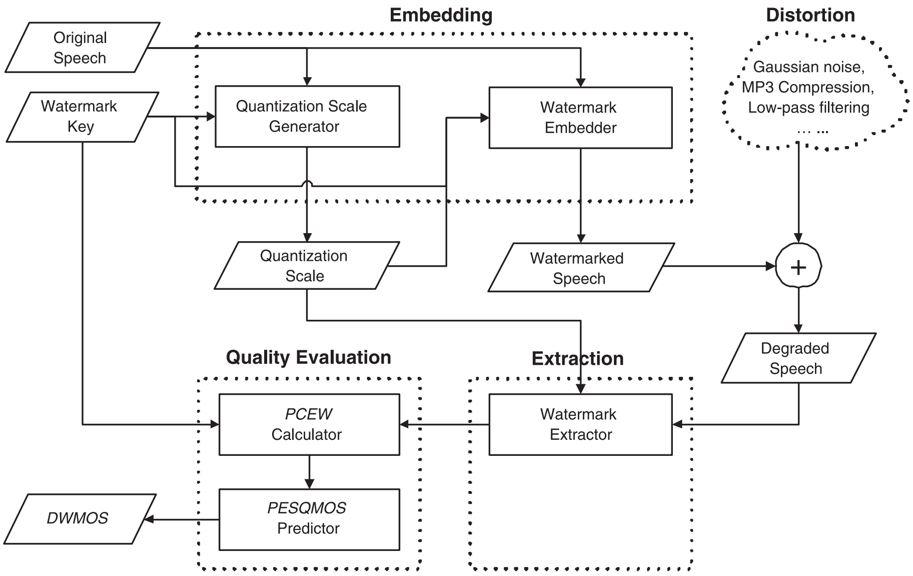


Fig. 1. Evaluation of speech quality using digital watermarking.

the selected coefficients, if $Q(e)$ is equal to the watermark bit (0 or 1), no change will be made to the coefficient, otherwise, it will be added a Δ to make the $Q(e)$ and the watermark bit match.

The watermark extraction is carried out in a similar procedure, and it only needs the Quantization Scale and watermarking key. By quantizing the DWT coefficients using Equ. (1), we can extract the watermark embedded.

2.2 Speech quality evaluation

After watermark extraction, the PCEW is calculated by comparing the extracted watermark with the original one based on Equ. (2).

$$PCEW = \frac{1}{N}\sum_{j=1}^{N} W(j) \oplus W^*(j) \tag{2}$$

where W is the original watermark, W^* is the extracted watermark, N is the length of the watermark, and $\oplus$ is the exclusive-OR operator. The PCEW value lies between 0 and 1.

We predict the speech quality from PCEW based on the mapping between ITU-T P.862 PESQ MOS (shortened as MOS in the rest of the paper) and PCEW. Hence, to give an accurate prediction, a linear mapping between MOS and PCEW must be calibrated. Through the experiments with twenty speech samples under ten different distortion parameters, it has been found that the mapping curves for the speech samples concentrate in a narrow band and are almost linear, and are independent of speakers.

In our method, we divide the mapping into 10 segments with a PCEW interval of 0.1, shown in Equ. (3).

$$\begin{cases} P_S = p & \text{if } p \leq PCEW < p + 0.1 \quad (p = 0, 0.1, ..., 0.9) \\ P_E = P_S + 0.1 & \end{cases} \tag{3}$$

DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004

where P_S and P_E are the percentages at the start and end point of the mapping segment, respectively.

And then, the PESQ score is predicted by the following equation:

$$DWMOS = MOS_S + \frac{MOS_E - MOS_S}{P_E - P_S} \times PCEW \tag{4}$$

where $DWMOS$ is the predicted MOS score using our digital watermarking based method; while MOS_S and MOS_E are the PESQ MOS values at the start and end point of the mapping segment.

2.3 Quantization step optimization

Different speech signals comprise different frequencies and amplitudes, therefore they have different robustness to the same distortion effect. DWT coefficients are real numbers. Their ranges vary and depend on both decomposition level and the speech itself. If we use different quantization step for watermarking different decomposition level of a speech, we will end up resulting in too many parameters sending to the watermark extractor. Hence, in our method, we introduce the term of Quantization Scale (QS) to obtain the quantization steps by using the following equation:

$$\Delta = \frac{\max_V - \min_V}{QS} \tag{5}$$

where Δ is quantization step, while $\max_V$ and $\min_V$ are respectively the maximum and minimum value of the coefficients in a specific decomposition level. We use the same QS for all the decomposition levels, however the quantization step Δ will normally be different because each level has different $\max_V$ and $\min_V$.

To obtain the optimal QS for each speech signal, we employ an adaptive control method to carry out recursive watermark embedding and extraction to make sure that the PCEW is in the range of $(0.995, 1)$ and the PESQ MOS is in the range of $(4.19, 4.21)$, before any distortion is applied.

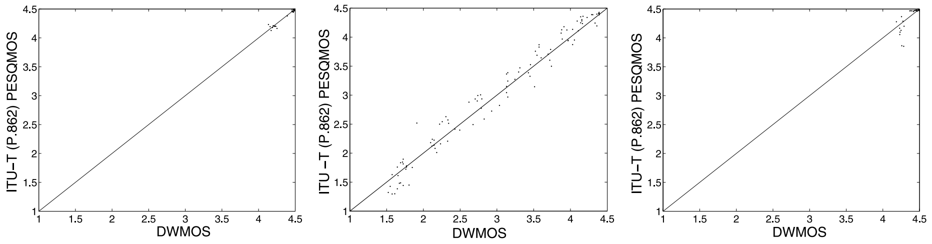
3 Experimental results and evaluations

We conduct the following experiments to demonstrate that the PCEW provides extremely high correlation with ITU-T P.862 PESQ MOS.

3.1 Sample speech selection

We selected two sets of samples that include both female and male speeches for different purposes. Set 1 was used for linear mapping calibration and Set 2 was for validation test. Both contain 10 speeches, which are stored in 16-bit, 16 KHz linear PCM format. For the distortions, we set SNR from 5 to 50 with an interval of 5 for Gaussian noise; bit rate from 32 to 320 Kbps with an interval of 32 Kbps for MP3 compression; and threshold frequency from 1 to 29 KHz with an interval of 4 KHz for low-pass filtering. Therefore, for the validation test, there are 100 speech samples for Gaussian noise and MP3 compression respectively, and 80 for low-pass filtering, as shown in Fig. 2.

DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004



(a) MP3 compression (100 samples)

(b) Gaussian noise (100 samples)

(c) Low-pass filtering (80 samples)

Fig. 2. Predicted MOS vs ITU-T P.862 PESQ MOS.

3.2 MP3 compression

For the MP3 compression, both our PCEW and the PESQ MOS curves show that when the bit rate is over 128 Kbps, the quality is approximately at the same excellent level. That means, the PCEW is approximately 100% and MOS is almost 4.5. When the bit rate is between 128 and 64 Kbps, the quality decrease linearly. When the bit rate is under 64 Kbps, the quality decreases very fast. The PCEW ranges from 0.42 to 1, while the MOS varies between 4.13 and 4.5. We predict the DWMOS values from PCEW using Equ. (4) in Section 2.2. The experimental results suggest that DWMOS and PESQ MOS have very close correlation (refer to Fig. 2 (a)), and that our quality evaluation has a pretty good accuracy on MP3 compression (refer to Section 3.5). Fig. 2 shows the correlation between the PESQ MOS and DWMOS. If the DWMOS and PWSQ MOS have an absolute match, all the sample points should be on the solid line from point $(1, 1)$ to point $(4.5, 4.5)$. The closer the sample points to the solid line, the better the performance of the DWMOS is.

3.3 Gaussian noise

For Gaussian noise addition, both the PCEW and PESQ MOS decrease with increasing noise strength. The MOS curves are more linear than the PCEW curves. However, the differences do not affect the accuracy of DWMOS because all the curves are very close and we can obtain a perfect mapping between PCEW and MOS. We predict the DWMOS values from PCEW using Equ. (4) in Section 2.2. As being indicated by Fig. 2 (b) and Table I, the DWMOS has very close correlation to the PESQ MOS, since all the sample points are distributed close to the solid straight line.

3.4 Low-pass filtering

Refer to Fig. 2 (c), there are 80 sample points, which were obtained from 10 test speeches with the threshold frequency from 1 to 29 KHz at an interval of 4 KHz. Under the low-pass filtering distortion, the PCEW curves are close and almost like straight lines with the values ranging from 0.2 to 0.99. Meanwhile, the PESQ MOS is not affected much by the low-pass filtering, with a lowest value round 4.07. When the threshold frequency is over 9 KHz, the PESQ MOS values are approximately the same and near 4.5. When the threshold frequency is below 5 KHz, the effect is bit more obvious. However,

DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004

because the the mapping curves between PCEW and PESQ MOS are very close, we can predict the MOS with extremely small errors, as shown in Fig. 2 (c) and Table I.

3.5 Accuracy of DWMOS

We use correlation coefficient and residual error between DWMOS and PESQ MOS to quantify the performance of our digital watermarking based speech quality evaluation method. Table I shows the results for MP3 compression, Gaussian noise addition, and low-pass filtering. In the table, "Set 1" was used for linear mapping calibration and "Set 2" was for validation test (refer to Section 3.1); "ARE" shorts for Absolute Residual Error; and "ARE $\leq$ C" means that the percentage of samples for which the ARE between the DWMOS and PESQ MOS is less than or equal to C. From Table I, we can see that our DWMOS is very well correlated to PESQ MOS.

Table I. Accuracy of DWMOS.

Distortion	Correlation coefficient		Mean	ARE	ARE	ARE
	Set 1	Set 2	ARE	$\leq$ 0.05	$\leq$ 0.25	$\leq$ 0.5
MP3 compression	0.9839	0.9727	0.0101	98%	100%	100%
Gaussian noise	0.9773	0.9759	0.1711	24%	85%	98%
Low-pass filtering	0.8501	0.8493	0.0361	80%	98 %	100%

4 Conclusion

In this paper, we proposed an objective speech quality evaluation method using digital audio watermarking based on Digital Wavelet Transform and quantization. The original speech signal is not needed, neither the training database. By comparing the DWMOS and ITU-T P.862 PESQ MOS values, we validated the accuracy of this method. Experimental results show that this method gives accurate predictions of subjective quality for speech signals. Furthermore, based on our experiments, our method can also be used for audio quality measurement.

DOI: 10.1587/elex.1.380
Received August 30, 2004
Accepted September 13, 2004
Published October 10, 2004