http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (3): 485-488, 2006
ISSN 1812-5638

A High-performance Psychoacoustics Approach to Speech Quality Evaluation

[1]Zhang Jun, [2]Gao Lei and [1]Zhang Deyun
[1]Network Laboratory, School of Electronics and Information Engineering,
Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, People's Republic of China
[2]Huawei-3com Technologies Co. Ltd, Beijing, 100085, People's Republic of China

Abstract: In order to solve the problem of real-time speech quality assessment, a high-performance algorithm, Efficient Psychoacoustics Evaluation of Speech Quality (EPESQ), based on psychoacoustics model, is proposed. The process of EPESQ is: The original and corresponding degraded speech samples are first preprocessed by overall gain compensation and IRS (Intermediate Reference System) filtering. Then both signals are transformed to their loudness presentation by a series of consecutive steps: windowed fast Fourier transform, frequency warping to Mel-scale and loudness mapping. The loudness presentations are compared in different time-frequency cell to get the differences called Disturbance. Disturbances are aggregated over time and frequency and then the result is processed by a cognitive formula to generate the final evaluation score. Experimental results show that EPESQ performs a 37.5% reduction in running time and 51.9% in memory occupation to the P.862 algorithm with only a 7.8% decrease in average correlation to listener opinions. EPESQ is a high-performance algorithm and suitable for real-time applications. It has been implemented in our Internet voice communication system as a self-evaluating component.

Key words: Speech quality assessment, pychoacoustics model, itensity warping, dsturbance aggregation

INTRODUCTION

Speech signal processing and transmitting techniques have made great progress in recent years. Various new speech communication systems have arisen. Evaluating the quality of speech signals after they are delivered by these systems becomes an important research area. The most reliable way to measure the quality of speech is to perform subjective speech quality assessment tests. In these tests, a large number of listeners listen to hundreds of short speech samples delivered by the system under test and rate its performance as a five-point scale. The average rating is expressed in Mean Opinion Score (MOS) proposed by International Telecommunication Union-Telecom Standardization Sector (ITU-T, 1996a). These tests are expensive, time-consuming and difficult to reproduce. Hence, it is desirable to have an objective method to simulate the subject rating procedure.

Several objective methods have been proposed (ITU-T, 1996c; Voran, 1999; ITU-T, 2001). The algorithm presented in the ITU-T Recommendation P.862 gains the highest precision among them. But it is too complex to be suitable for real-time scenarios (Hoene *et al.*, 2004). This paper proposes a psychoacoustics model based method, Efficient Psychoacoustics Evaluation of Speech Quality (EPESQ), which has much lower time complexity than the algorithm in P.862 with only a little decrease in precision.

SPEECH QUALITY EVALUATION BASED ON DISTANCE METRIC

Distance metric is a frequently adopted method in speech quality evaluation research. Its basic structure is shown in Fig. 1. The original speech signal and degraded signal are both preprocessed first, including IRS filtering, time alignment, level alignment, etc., (ITU-T, 2001). And then they are transformed by an auditory transform

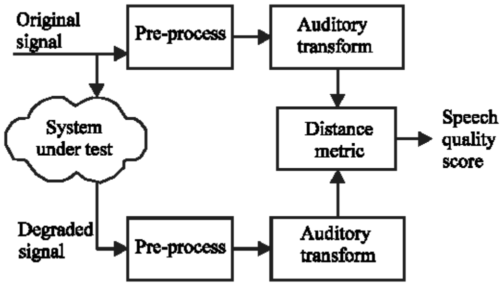


Fig. 1: Distance-metric-based speech quality evaluation

Corresponding Author: Zhang Jun, Network Laboratory, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, People's Republic of China Tel: 86-29-82668096 module to perceptual representations such as the frequency domain energy or the perceptual loudness. The transform results are fed into a distance-computing module to compare with each other in order to get the final speech quality evaluation score. Our approach also follows such a scheme.

EFFICIENT PSYCHOACOUSTICS EVALUATION OF SPEECH QUALITY

Our EPESQ algorithm, based on the psychoacoustics theory, is more efficient than the algorithm in P.862, with only a little decrease in precision.

Speech signal preprocess: Speech signal may be amplified/attenuated when passing through a transmitting system. The first step in the EPESQ algorithm is to rescale both signals to a correct overall level of 79 dB SPL (ITU-T, 1996b). We assume that variations between the levels of speech samples within a subjective test are very small and can be eliminated by setting the same overall level. Thus, we perform level alignment procedure to rescale the levels of the original and degraded signal over time and frequency (ITU-T, 2001). Ordinarily subjective listening tests are carried out using handsets. We adopt Intermediate Reference System (IRS) (ITU-T, 1989) to model the frequency response features of a handset. Speech signals often suffer from delay variations when passing through a packet-based transmission system. Hence, both signals need to be compared and aligned in time (Rix and Hollier, 2000) after filtered by IRS.

Psychoacoustics auditory transform: The auditory transform module in our approach is a psychoacoustics model. It consists of three consecutive parts: I time-to-frequency transform; II frequency warping; III loudness mapping. Detailed procedure is described below.

A Fast Fourier Transform (FFT) with a Hanning window is applied to calculate the power spectrum of each frame, which is 32 ms with 50% overlapping between two adjacent frames. The frequency is then mapped to Mel-scale (Davis and Mermelstein, 1980) linearly for those below 1000 Hz and using the following formula for the rest:

$$f_{mel} = 1000\log_2(1 + f/1000) \tag{1}$$

where, f_{mel} is the Mel-scale frequency and f is the linear frequency. The spectrum is then transformed to Mel Frequency Cepstrum Coefficients (MFCC) through a logarithmic Discrete Cosine Transform (DCT) after a filtering with 20 triangular band-pass filters. The Mel cepstrum is less in data amount and represents the acoustic information better than the Bark spectrum in P.862. Thus, it allows a significant computation reduction.

In order to reflect the human perceiving effects to speech signal, MFCC are mapped to loudness using the formula

$$L(f_{mel})_n = 0.715 \times \{10\log_{10}[P(f_{mel})_n/P_0(f_{mel})]\text{-}3.3\} \tag{2}$$

where, $L(f_{mel})_n$ and $P(f_{mel})_n$ are, respectively the loudness and the power of the speech signal in number f_{mel} Mel band of frame n and $P_0(f_{mel})$ is the hearing threshold which is set to different values for different Mel bands. Formula (2) is based on the psychology Weber's law that the human perceiving loudness is linearly in proportion to the logarithmic stimulus magnitude. The constant coefficients in (2) are derived from the regression analysis. The above mapping method is simpler in computation, hence more efficient in performance, than the Zwicker's law used in P.862.

Disturbance aggregation over frequency and time: To model how listeners compare the two signals, the difference between the original and the degraded signal loudness is calculated for each time-frequency cell (ITU-T, 2001). We call the absolute value of the difference disturbance. Disturbances of each time-frequency cell should be aggregated over frequency and time to generate a single value that characterizes the distortion, using L_p norm (ITU-T, 2001)

$$L_p = \left(\frac{1}{N}\sum_{n=1}^{N} D_n^p\right)^{1/P} \tag{3}$$

The disturbance is first aggregated over frequency across all time-frequency cells using a L_2 norm, resulting in a disturbance value for each frame. The L_2 norm for the frequency aggregation is

$$D_n = \sqrt{\frac{1}{k} A_{mel} \sum_{f_{mel}=1}^{k} \left(\left|D(f_{mel})_n\right| / P_n\right)^2} \tag{4}$$

where, D_n and P_n are, respectively the disturbance and the total power of the number n frame and $D(f_{mel})_n$ is the disturbance of number f_{mel} Mel-band in the n-th frame. A_{mel} is an ON/OFF weighting factor, equaling to 1 when P_n is below 60 dB SPL and 0 for the rest. This factor emphasizes the disturbance effect during relatively silent period, emulating the perceiving function of human auditory system.

The aggregation in time domain is performed next. Differing from the algorithm in P.862, which conducts a 2-phase aggregation, we adopt an overall L_1 aggregation for all frames:

$$D = \frac{1}{N}\sum_{n=1}^{N} D_n \tag{5}$$

Computation of evaluation score: The final step of EPESQ algorithm is to predict the subjective MOS score. We compute the score using

$$S = \alpha D^2 + \beta D + \gamma \tag{6}$$

where, S is the prediction of subjective MOS score and D is the disturbance generated by (5) and α, β and γ are constant coefficients, determined by the regression to the data getting from subjective tests. The regression was performed using a large amount speech file pairs. Each pair contains a degraded speech file to be evaluated and its corresponding original file of which MOS is available.

DISCUSSION

The precision of an objective speech quality evaluation method is determined by the correlation between the objective prediction and MOS. The correlation coefficient is calculated using

$$r = \frac{\sum(a_i - \bar{a})\sum(b_i - \bar{b})}{\sqrt{\sum(a_i - \bar{a})^2\sum(b_i - \bar{b})^2}} \tag{7}$$

where, a_i and b_i are, respectively MOS and the objective evaluation score under the number i test condition (noise, delay, code error, etc.) and $\bar{a}$ and $\bar{b}$ are the corresponding averages.

Using the 513 multilingual speech sample file pairs selected from ITU-T Coded-speech Database (ITU-T, 1998), we performed 11 evaluation tests in various conditions on a personal computer (Pentium 4, 512 MB). The different tests were conducted in a number of different languages. The precision of EPESQ is compared with the algorithm of P.862 and P.861 in Fig. 2 according to their correlations. Statistical results showed that the precision of EPESQ is 7.8% lower than P.862 algorithm and 11.6% higher than P.861 algorithm.

The time cost and space occupation results for the 513 file pairs, which were processed by EPESQ and the P.862 algorithm, respectively are shown in Fig. 3. Obviously, EPESQ has a much better performance in time cost and space occupation than the P.862 algorithm. On an average, EPESQ gives a reduction of 37.5% in running time and 51.9% in memory occupation to the P.862 algorithm.

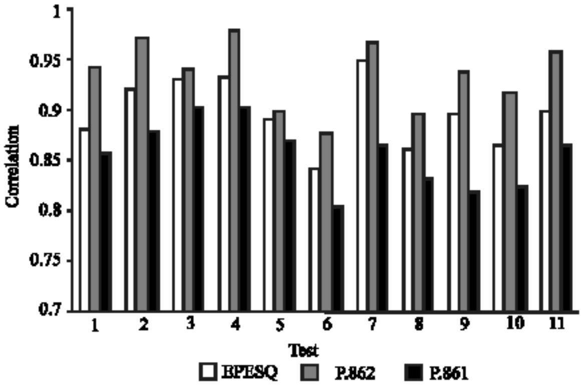
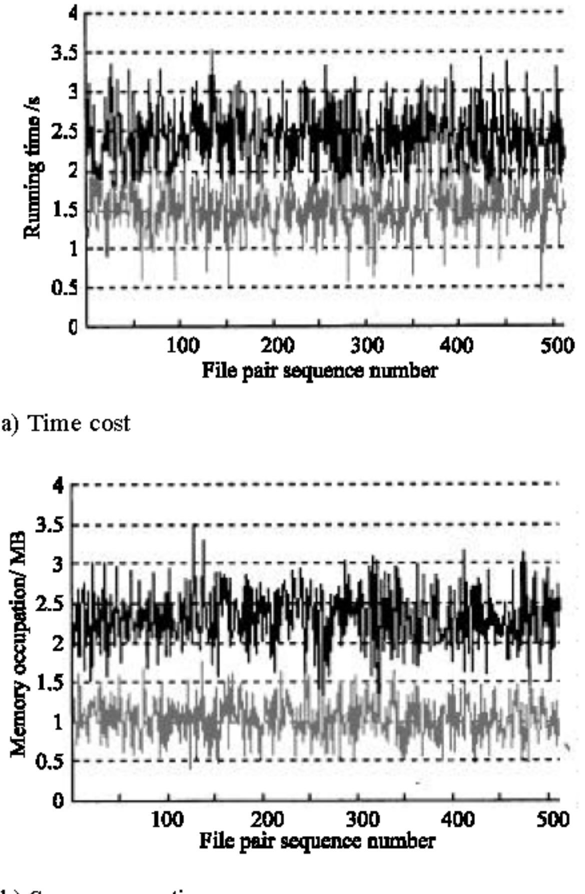


Fig. 2: Correlation coefficients comparison



(a) Time cost

(b) Space occupation

Fig. 3: Comparison of EPESQ and P.862 on time cost and space occupation

Figure 2 and 3 depict that EPESQ is more efficient than the algorithm of P.862 both on time and space, with only a little decrease on evaluation precision. So it is suitable for the scenarios of real-time applications and capability-limited micro systems.

In quality evaluation of telephone-band speech signal (300-3400 Hz), EPESQ performs more precisely than P.861 and other older models. It has a much lower time cost and space occupation than P.862 with only a little decrease in precision. It is greatly useful for real-time speech quality assessment and low cost embedded systems. We have applied it to the self-evaluation part of an Internet speech communication system to help assess speech quality and modulate transmission methods accordingly and EPESQ proved to be highly efficient and precise.

REFERENCES

Davis, S.B. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transaction on Acoustics, Speech and Signal Processing, 28: 357-366.

Hoene, C., H. Karl and A. Wolisz, 2004. A perceptual quality model for adaptive voip applications. International Symposium on Performance Evaluation of Computer and Telecommunication Systems, San Jose, California, USA.

ITU-T, 1989. Rec. P.48, Specification for an Intermediate Reference System.

ITU-T, 1996a. Rec. P.800, Methods for Subjective Determination of Transmission Quality.

ITU-T, 1996b. Rec. P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs.

ITU-T, 1996c. Rec. P.861, Objective Quality Measurement of Telephone-band Speech Codecs.

ITU-T, 1998. P-Series Supplement 23, ITU-T Coded-speech Database.

ITU-T, 2001. Rec. P.862, Perceptual Evaluation of Speech Quality (PESQ) -an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.

Rix, A.W. and M.P. Hollier, 2000. The Perceptual Analysis Measurement System for Robust End-to-end Speech Quality Assessment. IEEE International Conference on Acoustics, Speech and Signal Processing.

Voran, S., 1999. Objective Estimation of Perceived Speech Quality Part I: Development of the Measuring Normalizing Block Technique. IEEE Transaction on Speech and Audio Processing, 7: 371-382.