Further Investigations on the Relationship between Objective Measures of Speech Quality and Speech Recognition Rates in Noisy Environments

Francisco José Fraga[1], Carlos Alberto Ynoguti[2] & André Godoi Chiovato[2]

[1]Federal University of ABC – Engineering Center, Santo André - SP, Brazil
[2]National Institute of Telecommunications, Minas Gerais, Brazil
fraga@lsi.usp.br ynoguti@inatel.br agodoi@radial.br

Abstract

The relationship between an objective measure of speech quality (PESQ) and the recognition rate of a given speech recognition system was already investigated by other researchers. In this paper, we present a further investigation on such a relationship. In our research, the speech recognition tests were performed on a wider class of signals and SNR. The experimental setup as well as the speech recognition systems now evaluated were based on the directions given by the Aurora project. Moreover, a new parametric modeling approach for the PESQ-MOS versus speech recognition rate curve, based on the logistic function, is proposed. This new modeling allows some meaningful interpretations of the parameters of the logistic function in terms of system robustness, and permits to make inferences in the regions outside the experimental measures. Furthermore, the PESQ versus SNR characteristic was used to group types of noise, leading to a much better fit of the logistic function over the data points.

Index Terms: speech recognition, speech quality assessment

1. Introduction

Telephony speech signals are characterized by a great variability in the level and type of background noise, especially when arising from cell phones. In general, Automatic Speech Recognition (ASR) systems have a good performance in silent environments, but their performance used to drops dramatically in noisy environments. When these systems are used in call center applications, as an example, it would be useful if there was some kind of objective speech quality measure that could be able to predict the speech recognition rate of a given ASR system, without spending time and money to carry out extensive and expensive speech recognition tests.

Signal to noise ratio is the most common measure of a signal's quality, but normally it is not a good indicator of its *perceived quality* [1]. On the other hand, tests conducted by Sun and colleagues [2] suggest that the PESQ-MOS score, obtained by the Perceptual Evaluation of Speech Quality (PESQ) algorithm [3] is a better indicative for this purpose. In their work, they proposed an empirical 4th order polynomial fitting curve for modeling the relationship between the PESQ-MOS scores and the speech recognition rate, for some additive and convolutional noise scenarios.

This paper presents a further investigation of such a relationship between PESQ-MOS scores and the ASR rate, with three main contributions over Sun's et al. work [2]: 1) tests were performed on Aurora1 database, with a wider class of signals and SNR, providing a more robust evaluation; 2) a new parametric approach for the PESQ x recognition rate curve, based on the logistic function, is proposed, which allows some meaningful interpretations of the logistic function parameters in terms of the ASR system robustness; 3) the PESQ x SNR curve was used to group different types of noise that lead to similar ASR performance. These and other related issues are further discussed in the next sections of this article.

2. PESQ

When characterizing the perceptual quality of a speech signal, the Mean Opinion Score (MOS) is usually considered the most reliable test that can be performed. However, it is not practical on most cases due to the great human effort necessary to carry it out. The PESQ algorithm [3] was selected by the ITU to be a replacement of the subjective MOS evaluation in some well-defined situations.

The objective PESQ-MOS scores have a high correlation (more than 93%) with the subjective MOS scores under a wide range of conditions, and can be used to perform assessment of different codecs and end-to-end telecommunication networks. It takes two signals, a clean one and a noisy version of it, and compares both using perceptual models for the perceived pitch (Bark scale) and intensity (subjective loudness). There is a small scale shift between the MOS and the PESQ-MOS range: the MOS scale varies from 1 (worst quality) to 5 (best quality), whereas the PESQ scale range from –0.5 to 4.5.

3. PESQ x Recognition Rate Parametric Modeling

In [2], Sun et al. have tried to fit the PESQ x recognition rate data points by means of a 4th-order polynomial curve. But in our experiments, as we will show in the next sections, it was observed that the PESQ x recognition rate curve resembles the shape of a logistic function. Based on this similarity, we have proposed a parametric modeling of this curve using the logistic function. This new approach leads to some interesting interpretations, which never could be achieved by a non parametric polynomial curve, as shown next.

We used the logistic function with three free parameters: *a*, *b* and *c*, defined according to (1):

$$f(x) = c\left(\frac{1-e^{-ax}}{1+e^{b-ax}}\right)\times 100 \qquad (1)$$

Now, parameters *a*, *b* and *c* can be interpreted as follows:

3.1. Parameter *a*

This parameter controls the slope of the curve, as shown in Figure 1. It can be viewed as the "sensitivity of the system to PESQ variation". Observe that the higher is this parameter, the steepest is the curve.

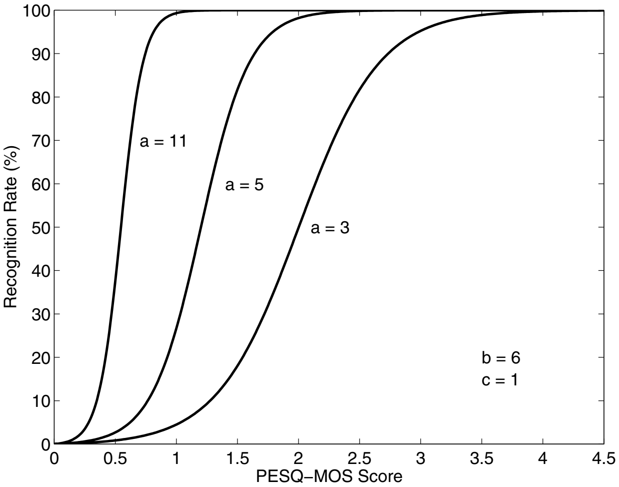

Figure 1 *Effects of parameter a in the logistic curve.*

3.2. Parameter *b*

Observing Figure 2, it can be easily noted that parameter *b* controls the horizontal offset of the curve: as *b* increases, the curve is shifted to the right. Analyzing this curve from right to left, it can be seem that, for fixed *a* and *c*, the lower is *b*, the better (in terms of robustness) is the ASR system, because it can keep high recognition rates for lower values of PESQ-MOS.

3.3. Parameter *c*

Parameter *c* controls the range of the curve in the vertical direction: as *c* increases, the recognition rate range increases too. This effect has an interesting interpretation: as the logistic function (1) tends monotonically to 1 as *x* tends to infinity, a value of $c = c_0$ will cause the curve to tend to c_0 as *x* tends to infinite. In our approach, *x* is the PESQ-MOS score, so when *x* is 4.5, it corresponds to the clean signal. Therefore, c_0 can be interpreted as the average recognition rate in clean conditions.

3.4. Merit index

We have also proposed an original *merit index* to rank a given speech recognition system regarding its robustness and recognition rate performance. In order to calculate the proposed *merit index*, it is necessary to establish an application-specific PESQ-MOS range. In next figures, all *merit indexes* were calculated over the 1.5 to 3.5 PESQ-MOS range. The *merit index*, which is a real number between 0.0 and 1.0, can thus be easily obtained by simple calculation of the logistic function mean value, using (1), except the factor of 100, with *x* sampled at the 0.5 distance points over the PESQ-MOS range. For example, in our above proposed PESQ-MOS range, *x* is sampled at the points $x_i = \{1.5, 2.0, 2.5, 3.0, 3.5\}$.

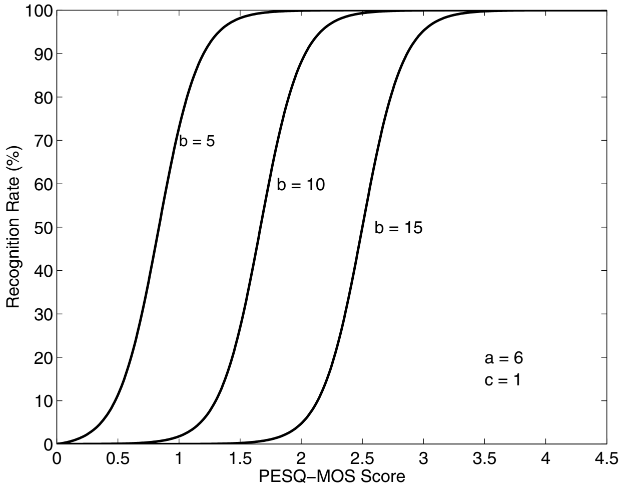

Figure 2 *Effects of b parameter in the logistic curve.*

4. Experimental Setup

The experimental tools used in this research consisted of a standardized speech recognition system, built using HTK [4] (for the back-end part of the system), and the Aurora1 database [5]. The whole word (ten digits plus *oh* and two "silent" models) HMM configurations used for the HTK back-end setup followed the detailed directions set for the first part of the Aurora project [5], which are, summarizing, the following ones: 16 states per word, simple left-to-right models (without skips over states), mixture of 3 Gaussians per state, diagonal covariance matrix.

Although well known, a brief description of the Aurora1 database is useful in order to show the dimension of the tests we have carried out. This database is derived from the TIDigits database, which is built by utterances from 110 adult speakers (55 male and 55 female) for training and 104 different adult speakers (52 male and 52 female) for tests, recorded in clean conditions (SNR > 30 dB). For each original utterance, eight types of noise were digitally added (subway, babble, car, exhibition hall, restaurant, street, airport and train station) at various signal-to-noise ratios (clean, 20dB, 15 dB, 10dB, 5 dB, 0 dB and –5dB). Four groups of 1001 original TIDigits (4004 original utterances) were used to form the noisy ones [5].

Furthermore, two different filters were used to simulate convolutional noise: G.712 (for conventional telephony) and MIRS (to simulate GSM-like channels) [6]. All the eight above mentioned noise types were filtered using G.712. Moreover, two additional types of additive noise were obtained by MIRS filtering the street and subway original noises, thus completing a total of ten types of noisy speech signals (Test A, B and C [5]), at seven different SNR, generating a total of 70,070 (10 noises x 7 SNR x 1001 original utterances) noisy signals for test.

The ASR system was trained using two distinct training strategies: clean training and multi-condition training; 8440 original TIDigits utterances were used in each training mode. The first one was performed using the filtered and 8 kHz downsampled TIDigits. The multi-condition training was accomplished by using a combination of clean and noisy speech signals; we followed exactly the detailed training and test setup explained in [5]. Two different ETSI STQ standardized front-ends were used for ASR performance evaluation: WI007 [5], which consists of log-energy and 12 mel-cepstrum coefficients plus their first and second derivatives, and the Advanced Front-End WI008 [7], which have state-of-the art noise reduction and blind equalization techniques (single channel) performed over the noisy signal before mel-cepstrum coefficients extraction.

The PESQ-MOS scores of each utterance of the test part of the Aurora1 database were calculated using as reference speech signals the clean filtered 8 kHz signals (PESQ = 4.5).

5. Results

Differently of that was observed by Sun et al. in [2], in a first attempt we did not get a good fitness of the PESQ-MOS x recognition rate data points by any single function, as can be observed in Figure 3. Here, there are 70 data points for each front-end (WI007 and WI008), corresponding to the ten types of noisy digits at seven different SNR. In this case, the word models were trained only by clean speech signals.

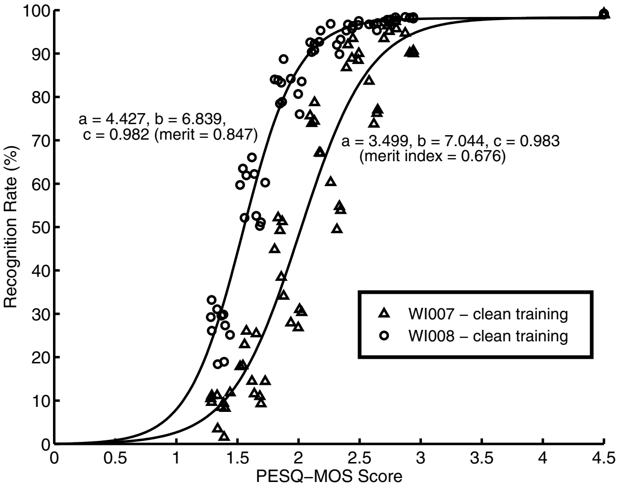

Figure 3 *Logistic function fit for WI007 and WI008*

When the ASR system was trained by a combination of clean and noisy speech signals (multi-condition training), the data points became less spread. This “concentration effect” can be clearly observed in Figure 4, where this time only the results when using the WI007 front-end are shown.

Although the fitting polynomial found by Sun and colleagues in [2] was obtained over a distinct database, with other noise types and a different ASR system, we also plotted in Figure 4 their fitting curve, just as a term of general comparison. Of course, any kind of strict comparison cannot be made between our results and the results related in [2], due the above mentioned reasons.

After have experimentally observed that it would be difficult to establish a general relationship between PESQ-MOS and speech recognition rate for all types of noise, in a second attempt we investigated if it was possible that such a strong relationship could exist for a limited group of noise types.

Following [2] once again, part of the experimental data analysis performed in this research was the observation of the relationship between PESQ-MOS scores and SNR for each type of noisy speech signal. Six of the total ten PESQ-MOS x SNR possible curves are plotted in Figure 5.

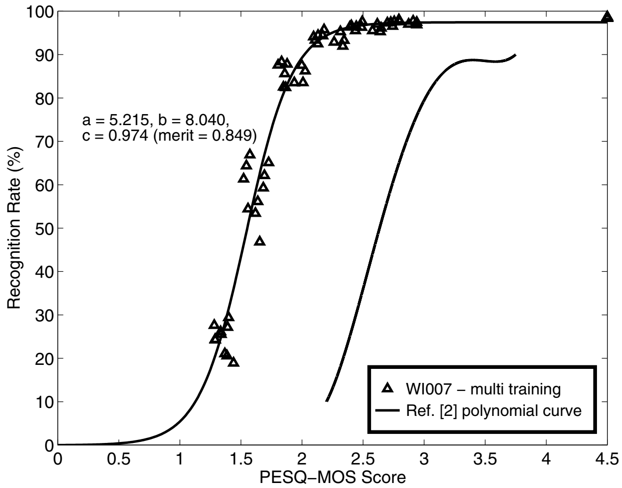

Figure 4 *Logistic function fit, WI007 – multi training*

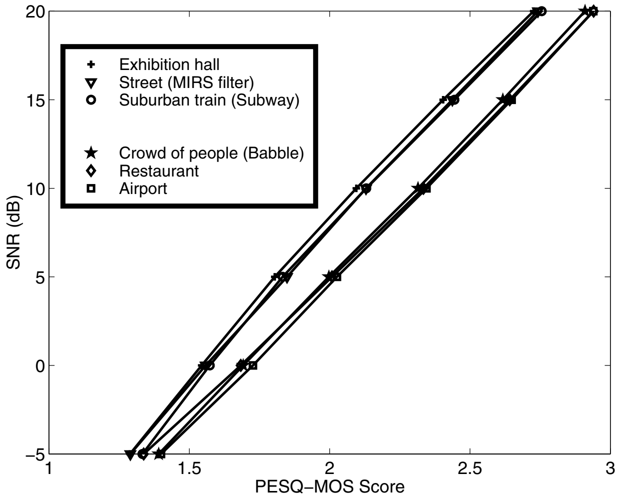

Figure 5 *PESQ-MOS versus SNR curves*

Besides confirming the almost linear behavior of these curves already reported in [2], we additionally observed that those noisy signal types could be clearly separated into two groups: the first group is formed by Exhibition hall, Street (MIRS filter) and Suburban train (Subway), while Crowd of people (Babble), Restaurant and Airport form the second group.

The first interesting result is that these two groups have also the same “group behavior” when we plotted their PESQ-MOS scores versus speech recognition rates, as undoubtedly shown by Figure 6. In this case, the ASR system was structured with a WI007 front-end and the “clean training mode” was used again.

Another surprising result is the fact that the noisy signals group from the right-hand side of Figure 5 has gotten a much inferior *merit index* than the other group, in spite of the fact that it presents a greater PESQ-MOS score than the left-hand side group for a given SNR. It means that, actually, it is not only the

SNR neither the speech quality alone, measured by means of the PESQ algorithm, that governs the speech recognition performance, but a combination of both. Given two groups of noisy digits (obtained from the same clean signal but with different types of additive noise), each one with the same mean PESQ-MOS scores, the group presenting greater SNR will present a much better performance. This can be clearly observed in Figure 6, at PESQ-MOS Score near to 2.0: a difference of almost 50% (30% versus 80%) on Recognition Rate was achieved by the second group of noisy signals over the first one.

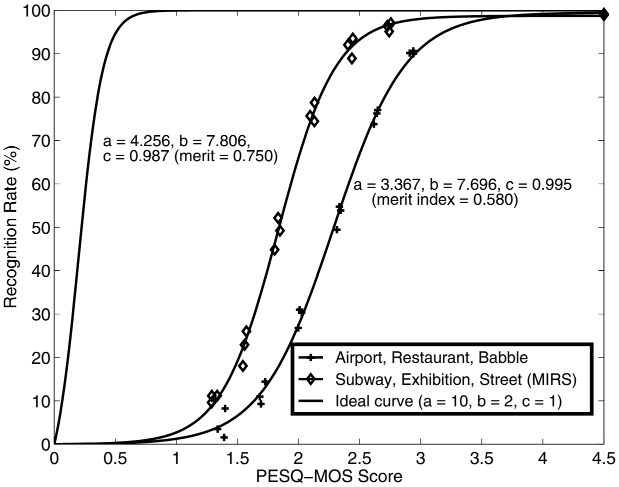

Figure 6 *Group Logistic functions, WI007 – clean training*

When the noisy signals are grouped according to their position in the PESQ-MOS x SNR plot, a strong relationship between PESQ-MOS and speech recognition rate can now be observed, and the logistic curve fits (using the MSE criterion) the data points with great success, as we can see in Figure 6.

Finishing our discussion, the following statements can be established as a consequence of the above related results:

I. Based on our 70-point plots (for each ASR system setup), it was not possible to get a good fit of the PESQ-MOS x recognition rate points by means of any single function;

II. On the other hand, a very good fit (in the MSE sense) of the PESQ-MOS x recognition rate data points can be done by the logistic function when the noisy speech signals are grouped according to a good criterion;

III. The PESQ-MOS x SNR characteristic curve of each type of noisy signals is a good criterion to perform their group classification, leading to a very good "logistic type" of PESQ-MOS x recognition rate curve behavior when this grouping criterion is used;

IV. After adequately grouped into noisy signals type classes, a given ASR system in a given "environment class" can be ranked by means of a single number, called *merit index*, which should be calculated using the logistic function parameters over a pre-determined and application-specific PESQ-MOS range;

V. The well-defined behavior of the logistic function permits to predict the ASR rates outside the experimental obtained values; this characteristic is not achievable by any polynomial fitting curve.

6. Conclusions

An extensive investigation regarding the relationship between automatic speech recognition rates (gotten by a standardized ASR system under several noisy conditions [5]) and the ITU-T P.862 PESQ-MOS has been performed. Unlike reported in [2], a general relationship between PESQ-MOS speech quality measures and recognition rates for all noise types could not be well established.

On the other hand, a very good fit of the PESQ-MOS x recognition rate relationship can be done by a single logistic function when the noisy speech signals are grouped according to its PESQ-MOS x SNR curve. Furthermore, such a modeling allows some meaningful interpretations on the logistic function parameters in terms of the ASR error rate, and, unlike polynomial curves, it permits to make inferences in the regions outside the experimental measures.

An important consequence is that the PESQ-MOS may be used to predict the average achievable speech recognition rates in real-life applications, under certain conditions. Moreover, a *merit index* based on the fitted logistic curve was proposed to rank a given ASR system in a given group of ASR application scenarios.

7. Acknowledgements

We would like to thank for FINATEL and FAPESP the partial funding of this research.

8. References

[1] Hansen, J. H. L. and Pellom. "An effective quality evaluation protocol for speech enhancement algorithms". In Proceedings of the ICSLP'1998, Sydney, Australia, pp 2819-2822, 1998.

[2] Sun, H., Shue, L. and Chen, J. "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech". In Proceedings of the ICASSP'2004. Montreal, Canada, pp. 865-868, 2004

[3] Beerends J., G., Rix A.W., Hollier M. P. And Hekstra A. P., "Perceptual Evaluation of Speech Quality (PESQ). The New ITU Standard for End-to-End Speech Quality Assessment'". Journal of Audio Eng. Soc., vol. 50, no. 10 pp., 755-778, 2002.

[4] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell J., Ollason, D., Valtchev, V. and Woodland, P., "HTK Book (for HTK Version 3.1)". Cambrigde University Engineering Departament, 2001.

[5] H.G. Hirsch, D. Pearce. "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, Paris, France, September 18–20, 2000.

[6] ETSI-SMG Technical Specification. "European Digital Cellular Telecommuncation System (Phase 1), Tranmission Planning Aspects for the Speech Service in GSM PLMN System". GSM03.50, version 3.4.0, 1994.

[7] ETSI ES 202 050 V1.1.3 (2003-11) ETSI Standard. "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2003.