Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification

Alberto Harriero, Daniel Ramos, Joaquin Gonzalez-Rodriguez, and Julian Fierrez

ATVS – Biometric Recognition Group,
Escuela Politecnica Superior, Universidad Autonoma de Madrid,
C. Francisco Tomás y Valiente 11, 28049 Madrid, Spain
{alberto.harriero,daniel.ramos,joaquin.gonzalez,
julian.fierrez}@uam.es

Abstract. In this work, we analyze several quality measures for speaker verification from the point of view of their utility, i.e., their ability to predict performance in an authentication task. We select several quality measures derived from classic indicators of speech degradation, namely ITU P.563 estimator of subjective quality, signal to noise ratio and kurtosis of linear predictive coefficients. Moreover, we propose a novel quality measure derived from what we have called Universal Background Model Likelihood (UBML), which indicates the degradation of a speech utterance in terms of its divergence with respect to a given universal model. Utility of quality measures is evaluated following the protocols and databases of NIST Speaker Recognition Evaluation (SRE) 2006 and 2008 (telephone-only subset), and ultimately by means of error-vs.-rejection plots as recommended by NIST. Results presented in this study show significant utility for all the quality measures analyzed, and also a moderate decorrelation among them.

Keywords: Speaker verification, quality, utility, SNR, degradation indicator.

1 Introduction

Speaker recognition is nowadays a mature field with multiple applications in security, access control, intelligence and forensics. The State of the Art is based on the use of spectral information of the speech signal, combining such information in multiple ways, and compensating the inter-session variability of speech recordings [1,2].

Despite the significant advance on the performance of the technology in the field, partly due to the efforts of NIST and their successful periodic Speaker Recognition Evaluations [3], the field of speaker recognition faces important challenges. Among them, performance of comparisons when there is a high mismatch between enrollment and testing speech conditions is far from being solved, although the improvements in this sense in the last years have been remarkable [1]. Moreover, the mismatch in the conditions of the speech databases for system tuning and for operational work (the so-called database mismatch problem [4]) has a strong impact in the performance of the systems, and attenuates the beneficial effects of compensation techniques.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 434–442, 2009.

In order to solve the problems associated to session variability in speech, the speaker recognition scientific community continues their efforts on improving the existing compensation algorithms [1]. These methods are mainly based on data-driven approaches modeled with statistical techniques such as factor analysis [1]. Although their demonstrated success, such techniques are sensitive to the existence of a rich development corpus, desirably in similar conditions to those of the operational framework, which may not be available in general. Moreover, there is other knowledge about the speech signal which can be efficiently extracted from excerpts and used as information about the variability of the speech signal and its impact on the performance of speaker recognition systems. Among such knowledge are the quality measures, as recently proposed by NIST [5].

In this work, we present an analysis of several quality measures from the point of view of their utility, i.e., their usefulness as a predictor of system performance. Some of the analyzed quality measures are derived from classical indicators of speech degradation, namely Signal to Noise Ratio (SNR), statistics from Linear Predictive Coefficients (LPC) and estimators of subjective quality (such as ITU P.563 recommendation [6]). Moreover, we propose a quality measure with an attractive interpretation, derived from what we have called Universal Background Model Likelihood (UBML). The work also presents a framework for the obtaining of the proposed quality measures from speech. The paper is completed with experimental results using telephone speech and protocols from recent NIST Speaker Recognition Evaluation Evaluations (SRE), where the utility of quality measures is shown by the performance measures recommended by NIST [5].

The paper is organized as follows. In section 2, we define the quality measurement framework according to previous work in the literature [6,7], We also present three classical quality measures derived from classical indicators of speech degradation. In section 3, we present a novel quality measure based on what we have called the Universal Background Model Likelihood (UBML). Results showing the analysis of the four analyzed quality measures, including the proposed one derived from UBML, are described in section 4, where the utility of the proposed measures is analyzed using two different databases from NIST Speaker Recognition Evaluations (2006 and 2008). Experiments allow the identification of the most useful quality measures for predicting performance, based on protocols recommended by NIST [5]. Finally, conclusions are drawn in section 5.

2 Quality Measures for Speaker Verification

The idea that the quality of the speech signal affects the ability of an automatic system to distinguish among people from their voices is somewhat intuitive, as it happens in other biometric traits [8]. In fact, the measurement of speech quality has been a major topic of research during the last decades [9]. The need to monitor the quality of speech signals on telephone networks has lead to the development of several algorithms to estimate the subjective quality of a speech signal [9], understood as the quality perceived by a given user. The recommendation P.563 of the International

[1] The last research workshop on the topic at John Hoskins University deserves special attention (http://www.clsp.jhu.edu/workshops/ws08/groups/rsrovc/).

Telecommunications Union (ITU) [6] is an estimation method of the subjective speech quality which includes the effects of the majority of existing impairments in modern telephony networks. Its output is computed from 51 parameters, which are indicators of different possible degradations. The quality measures from this study are mainly based on degradation indicators found in ITU P.563 as well as other work in the literature [10].

According to previous work in the literature [6,8], we define a quality measure as a scalar magnitude which predicts the performance of a given biometric system. Under such a definition, utterances with poor quality are more likely to be misclassified than those of good quality. A quality measure is defined to be bounded in the range between 0 and 1, where 0 corresponds to the worst possible quality value and 1 to the best one. As this scalar is based on parameters which, in general, do not belong to this range, a mapping function has to be applied, in such a way that for every possible value of a degradation indicator x, the mapping assigns a quality value $\mathrm{Q(x)} \in [0,1]$.

The evaluation of quality measures is carried out following the recommendations given by NIST [5], according to which a quality measure is considered useful if as we reject scores with the lowest quality values, the system performance improves.

2.1 Classical Quality Measures

Quality measures defined in this section have been used before with the purpose of evaluating speech degradation [6,10].

Signal to Noise Ratio (SNR). The SNR degradation indicator has been calculated as follows: making use of a energy-based voice activity detector, each utterance is separated in non overlapping voiced and un-voiced frames of 20 ms. Then, average energy is calculated for both types of frames. Finally, SNR is computed as:

$$SNR = 10 \times \log\left(\frac{E_{voiced}}{E_{unvoiced}}\right). \tag{1}$$

where E_{voiced} and $E_{unvoiced}$ are the mean energies of the voiced and unvoiced sections. This method for measuring SNR has one main drawback: as it depends on the VAD accuracy, it may have problems to differentiate voiced from un-voiced sections for noisy or very high activity utterances.

We defined the SNR quality mapping function as:

$$Q_{SNR}(x) = \frac{x}{60}. \tag{2}$$

where x is the SNR value, which is supposed to belong to the range 0-60 dB. Values outside this range will be limited prior to mapping to quality.

Kurtosis LPC (KLPC). Kurtosis is a 4th order statistic which measures the degree of fat tails of a distribution. In this case, kurtosis is applied to the LPC coefficients distribution, as is done in ITU P.563 recommendation [6]. For every 20 ms frame, 21 LPC coefficients are obtained. Then, kurtosis is calculated as:

$$k = \frac{1}{P}\sum_{p=1}^{P}\left(\frac{a_p - \frac{1}{P}\sum_{p=1}^{P} a_p}{\sigma}\right)^4. \tag{3}$$

where σ represents the standard deviation of LPC coefficients, a_p. Finally, all kurtosis values from all the voiced frames are averaged.

As it will be shown later, the system performance decreases as KLPC increases. According to this, we defined its mapping function as:

$$Q_{KLPC}(x) = 1 - \left(\frac{x-3}{8}\right) . \tag{4}$$

where x is the KLPC value, which based on our experiments, is supposed to belong to the range 3-11.

ITU P.563 Recommendation (P.563). ITU provides an implementation of the algorithm defined on this recommendation. The algorithm generates a Mean Opinion Score (MOS) [11] for each utterance, which is representative of the utterance subjective quality. The MOS belongs to the range 1-5, where 1 corresponds to the worst possible quality value, and 5 to the best one. The input utterance must have a length between 3 and 20 seconds. All utterances duration were between 2 and 5 minutes long, so they had to be divided in smaller fragments and their MOSs were averaged.

The mapping function has been defined according to the MOS scale:

$$Q_{P563}(x) = \frac{(x-1)}{4} . \tag{5}$$

3 UBML: A Novel Quality Measure for Speaker Verification

In this work we propose a degradation indicator in the context of speaker verification based on Gaussian Mixture Models (GMM) [13], although the approach can be used in any possible system, no matter the modeling scheme. The proposed measure is motivated by a simple idea. Given that a Universal Background Model (UBM) from a GMM represents the common distribution of speaker features for a given expected operational database, degraded signals are more likely to differ from a UBM than non-degraded signals. Thus, the likelihood between any utterance and the UBM can be used as a measure of speech degradation. Moreover, it is well-known that speech utterances not matching a given UBM in a GMM system will tend to perform poorly, and therefore a simple measure of the match between a given speech utterance and the UBM like UBML will predict performance for any utterance. Although it may be argued that the likelihood with respect to a UBM may represent many other speaker-dependent information non related to speech degradation, experiments with UBML showed a strong relationship between system performance and this indicator, supporting the assumed hypothesis. In section 5, the validity of this measure is further discussed.

Obtaining UBM likelihood is a mandatory step when using a GMM system, and therefore if such a system is used, the obtention of UBML indicators is costless. However, for other systems UBML can be previously computed and its quality measure used as well. Given a speaker GMM model λ_t and any utterance O for which feature vectors have been extracted, a similarity score is typically computed as:

$$S(O, \lambda_t) = \log\, p(O, \lambda_t) - \log\, p(O, \lambda_{UBM}) . \tag{6}$$

where $p(.,\lambda)$ is the probability density function for any model λ. The last term gives the likelihood between any utterance and the UBM:

$$\mathrm{UBML} = \log p(O, \lambda_{UBM})\ . \tag{7}$$

We define the mapping function based on the typical distribution of UBML according to the experiments performed in this work, whose values lay within the range (-13,-5). It is expected that for a given GMM system configuration this value will not significantly change its range among databases. Thus, we map the quality measure as follows:

$$Q_{UBML}(x) = \frac{(x+13)}{8}\ . \tag{8}$$

4 Experiments

4.1 Databases, Systems and Protocols

In order to evaluate the utility of quality measures, we have used telephone databases and protocols from NIST Speaker Recognition Evaluations 2006 and 2008, which represents a real challenge in terms on session variability [3]. We have selected both corpuses for experiments in order to show the general behavior of the proposed quality measures among different telephone databases. This fact allows a general strategy of training quality mappings from degradation indicators using a given database (namely NIST SRE 2006) and using such mapping on a different one (namely NIST SRE 2008). For NIST SRE 2008, we have selected the telephone-only subtask of the core condition, namely *short2- short3 tlf-tlf*. For NIST SRE 2006, the whole core condition is used, namely *1conv4w-1conv4w*. For both conditions in the different evaluations, each conversation (coined *short2* for training and *short3* for testing) has an average duration of 5 minutes, with 2.5 minutes of speech on average after silence removal. Variability due to different transmission channels, languages and environmental conditions is present, but even more accused in SRE 2008. Although there are speakers of both genders in the corpus, no cross-gender trials are defined.

For score computation, the ATVS GMM system has been used, where speech data known to come from a given speaker is represented using Gaussian Mixture Models adapted from a Universal Background Model. The front-end consists of the extraction of 19 MFCC plus deltas, and processed with rasta filtering and feature warping. Channel factors at feature level have been used for channel compensation [1]. GMM of 1024 mixtures have been used for modeling. Finally, T-Norm has been used for score normalization. The background set for T-Norm cohorts, channel compensation and background modeling is a subset of databases from previous NIST SRE.

4.2 Degradation Indicators Evaluation

Experiments presented in this section were carried out for 12 different degradation indicators, from previous work in the literature [6,7,10] . They were intended to show the variations of the system performance depending on the magnitude of each indicator, which is useful in order to determine the mapping function from indicator to quality measure. From the whole set of 12, we selected those which showed a clearer relationship with the system performance, namely SNR, ITU P.563 and KLPC.

The experiment was carried out as follows:

1. For every utterance in the databases, each degradation indicator was computed.
2. Scores from the experimental set-up were computed for the described protocols and using the ATVS GMM system.
3. For every score *i*, a mean degradation indicator μ_i is generated computing the arithmetic mean of the indicators for the training utterance and the test segment.
4. Scores are arranged according to their mean degradation indicator μ_i.
5. The first 20% of ordered scores are selected. This is known as set *k*. For each score set *k*, the EER$_k$ is computed, as well as the mean degradation indicator.
6. The last step is repeated 100 times for each set of scores *k*=1,…100. Each time selecting a set of scores with higher degradation indicator. The last set will correspond to the 20% scores with highest degradation indicator.
7. As a result, we obtain 100 EER values and 100 mean degradation values, which correspond to 100 overlapped sets of scores. EER is then represented with respect to its corresponding mean set degradation value.

The following plots show the result of the best performed degradation indicators from the 12 analyzed. We also show the results for the proposed UBML.

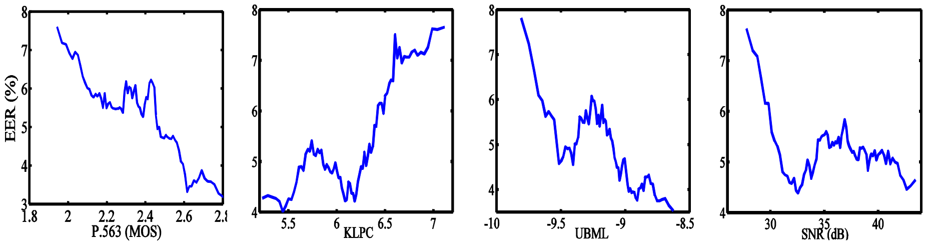

Fig. 1. EER (%) for every set of scores with a given mean indicator value: P563, KLPC, UBML and SNR, for the NIST SRE 2006 database

As we can observe, all of them show a clear relationship with the system performance, particularly UBML and P563, for which the EER decreases roughly from 8% to 4% for the set of scores with highest qualities.

4.3 Correlation Experiments

Given any two quality measures, the linear correlation coefficient among them gives an estimate of how similar is the information they provide about speech degradation in each utterance. This may be interesting in order to combine different quality measures and to optimize the available information to discriminate degraded quality samples. On the following tables we show the correlation coefficients for the five quality measures for both SRE 2006 and 2008 databases.

As we can observe, in general all correlation values are moderate. It can be observed a remarkable correlation between UBML and the measures P.563 and SNR. Since P.563 and SNR are well-known degradation indicators, this fact confirms the hypothesis stated in Section 4: UBML is an indicator of signal degradation.

Table 1. Correlation coefficients for the four quality measures: snr, klpc, p563 and ubml

	SRE 2006			SRE 2008		
	snr	klpc	ubml	snr	klpc	ubml
p563	0.136	0.192	0.223	-0.005	0.145	0.097
snr		0.182	0.386		-0.132	0.536
klpc			-0.034			-0.281

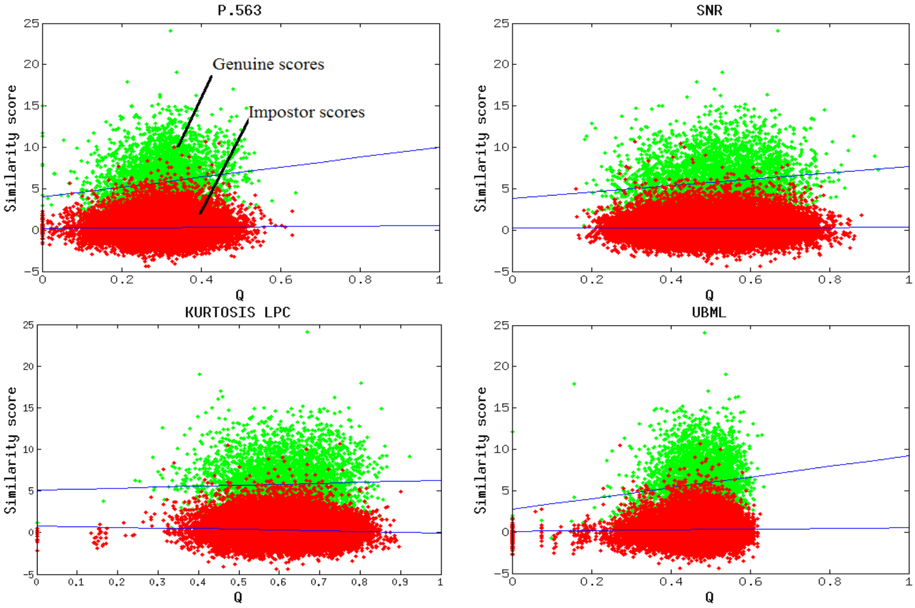

Fig. 2. Similarity scores against Q, for every quality measure for the SRE 2008 database

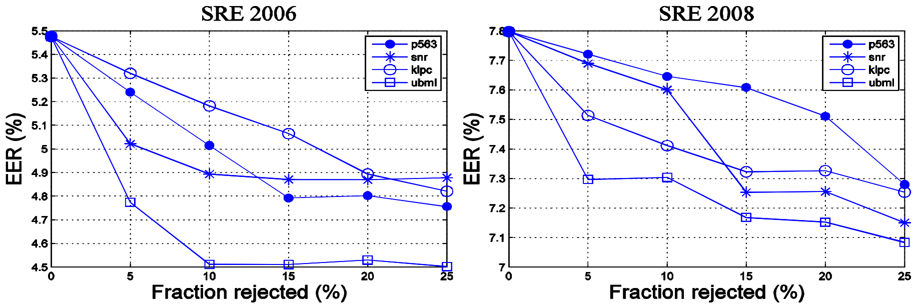

Fig. 3. EER (%) against rejected scores (%), for both NIST SRE 2006 and 2008 databases

SNR measure presents a low correlation with P.563. This may be due to the low noise level of both databases, since P.563, which selects the strongest of several degradation indicators, is not considering SNR a dominant one. However, SNR has a clear correlation with UBML, which means that the likelihood between any utterance and the UBM is quite sensitive to the noise contained in the utterance.

4.4 Utility Experiments

In this section we try to show the effectiveness of the quality measures as predictors of the system performance. We make use of two kinds of graphic representations: scores-vs-quality scatter plots and error-vs-rejection plots. On the first one, we represent the similarity score against their corresponding quality values (Q), which are obtained combining the qualities of the two involved utterances as:

$$Q = \sqrt{Q_t \cdot Q_{tr}}. \qquad (9)$$

where Q_t and Q_{tr} are the quality measures of the test and train utterances.

Since better quality values are supposed to predict better results, target and non-target scores should get more separated as Q gets more close to 1. Regression lines fitted on the plots are intended to show this tendency.

As we can observe, for the quality measures P563, SNR and UBML, scores show a clear tendency to get separated for higher values of Q.

Finally, EER vs reject plots are used as recommended by NIST to show the utility of quality measures [5]. In these plots, the EER is represented against a given percentage of scores rejected with lowest quality values. The curve is supposed to decrease if the quality measure is useful as the rejection percentage increases. We have represented the results for the rejection fractions: 5, 10, 15, 20 and 25%.

We can observe that EER decreases for all the quality algorithms as we reject scores. In general, all measures perform better for the 2006 database. It is worth noting that UBML is the best performed measure for both databases, especially for 2006, where the EER decreases a 20% after rejecting the 10% of the scores.

5 Conclusions

In this paper we have analyzed the utility of several quality measures obtained from different indicators of speech degradation typically used in speech processing, namely ITU P.563 estimator of subjective quality, signal to noise ratio (SNR) and LPC Kurtosis (KLPC). We have also proposed a novel quality measure based on the likelihood of a speech segment with respect to a universal model (UBML), which measures degradation in a speech segment by its divergence with respect to such a model. Performance of the quality measures has been presented following the recommendations by NIST, and also using different databases and protocols from NIST Speaker Recognition Evaluations. In all cases, a remarkable utility has been obtained, and a moderate correlation has been observed among different quality measures. Thus, we can argue that the analyzed measures are predictors of speaker verification performance, and therefore they can be used as information in order to compensate for performance drops due to speech degradation.

Future work is mainly related with the use of the obtained quality measures for improving speaker verification performance, and also as complementary information to other data-driven approaches for session variability compensation or fusion in speaker recognition. The potential uses of the promising UBML-based quality measure will be also explored in depth. Finally, a more complete classification of quality measures for speaker verification will be also addressed, including the utility analysis of other different quality measures.

Acknowledgements

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. The authors thank Fernando Alonso-Fernandez and Ignacio Lopez-Moreno for fruitful discussions and suggestions.

References

1. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing 16(5), 980–988 (2008)
2. Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D., Matějka, P., Schwarz, P., Strasheim, A.: Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Transactions on Audio, Speech and Language Processing 15(7), 2072–2084 (2007)
3. Przybocki, M.A., Martin, A.F., Le, A.N.: NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006. IEEE Transactions on Audio, Speech and Language Processing 15(7), 1951–1959 (2007)
4. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J.J.: Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. In: Proc. Interspeech 2008, vol. 1, pp. 1493–1496 (2008)
5. Grother, P., Tabassi, E.: Performance of Biometric Quality Measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 531–543 (2007)
6. Malfait, L., Berger, J., Kastner, M.: P.563-The ITU-T Standard for Single-Ended Speech Quality Assessment. IEEE Trans. On audio, speech and language processing 14(6)
7. Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: Using Quality Measures for Multilevel Speaker Recognition. Computer Speech and Language 20(2,3), 192–209 (2006)
8. Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Fronthaler, H., Kollreider, K., Bigun, J.: A comparative study of fingerprint image-quality estimation methods. IEEE Trans. on Information Forensics and Security 2(4), 734–743 (2007)
9. Grancharov, V., Kleijn, W.B.: Speech Quality Assessment. Springer Handbook of Speech Processing. Springer, Heidelberg (2008)
10. Richiardi, J., Drygajlo, A.: Evaluation of speech quality measures for the purpose of speaker verification. In: Proc. of Odyssey 2008, the ISCA Speaker and Language Recognition Workshop, Stellenbosch, South Africa (2008)
11. Mean opinion score (MOS) terminology, ITU-T Rec. P.800.1 (2003)
12. Reynolds, D.A.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)