Evaluation of speech quality measures for the purpose of speaker verification

Jonas Richiardi, Andrzej Drygajlo

Signal Processing Institute
Swiss Federal Instiute of Technology (EPFL)
jonas.richiardi@epfl.ch

Abstract

Real-world deployment of speaker verification systems often have to contend with degraded signal quality and erratic statistical behaviour of the speech data being modelled. We present signal quality estimation techniques for extraction of additional information about the speech data that can be used to improve performance of speaker verification systems in degraded conditions. We propose methods to perform objective evaluation of these quality measures for the purpose of their comparison using benchmarking databases, and show why the class must be taken into account when evaluating quality measures.

1. Introduction

Many factors conspire to cause verification errors in speaker verification. Variability in acquisition conditions, as well as variability of the users' presentations entail a certain level of uncertainty in the classifier's decision. In order to address these issues and improve classification performance, it is crucial to be able to measure phenomena which may be indicative of variability.

A *quality measure* is a measurable indicator of a factor impacting the classifier behaviour, which exhibits a dependency relationship with the classifier output scores and/or classifier decisions. It can be jointly modelled with the classifier's scores or decisions in order to improve the verification result or provide estimates of the reliability of the verification result.

In pattern recognition terms, quality measures constitute features. They are used in single-classifier systems, where they are crucial because they provide additional information which can help a meta-classifier to improve upon the results of both the base classifier and a meta-classifier using only scores or decisions. They are also used in multiple-classifier systems, where they help explain the relationships between classifiers, leading to classifier combination models that outperform combination models using only the hard or soft output of classifiers.

In speaker verification, degraded acquisition conditions resulting in additive noise or channel noise cause utterance-dependent errors. There are many approaches to handle mismatch and afford robustness to the classification, many inspired by similar work in speeech recognition, at all level of the pattern recognition chain:

One approach is feature compensation, whereby features are transformed using some warping function to conform to an expected distribution[1, 2].

The most commonly used approach consists in normalising the score obtained on the user model by the score obtained on the background model; the idea being that the condition mismatch will affect both models and thus compensate for user model score drift [3, 4].

Another possibility is to have explicit models of scores under certain degraded conditions, possibly incorporating quality measures [5, 6]. This is the approach we favour, as if offers intepretability and it has been reported that filtering or compensation approaches afford only limited robustness [7]. Furthermore, it is usable in conjunction with other robustness methods.

In all cases where the signal quality is explicitely modelled (such as [8]), quality measures play an essential role. In order to improve on verification results, it is important to develop signal quality measures which have a dependency relationship with the classifier output. This paper proposes to explore some of the issues related to the use of quality measures in speaker verification.

In Section 2, we propose a method for evaluation of quality measures. We review speech quality measures in Section 3, and conclude with experimental evaluation of quality measures in Section 4.

2. Evaluating quality measures

2.1. Visual inspection

Since one possible use of quality measures is to predict verification errors, a way of evaluating quality measures is to plot their distributions with respect to two classes: the class of correct classification decisions, and the class of incorrect classifications, which we denote Decision Reliable: $DR = 1$, respectively $DR = 0$. These densities can be obtained in several ways, but we recommend kernel-based density estimation, histograms or mixture models because many times these distributions will be asymmetrical and multimodal.

2.2. Assuming homoscedasticity of scores

A simplifying assumption that can be made is that the variance of the score is equivalent throughout its range. While this does not hold in practice, it allows for the definition of simple measures of performance for quality measures.

2.2.1. Assuming linearity of relationships with quality measures

Quality measures can be evaluated by measuring their statistical dependence on the scores. Under the assumption of linearity this dependence can be estimated by computing the correlation coefficient between the quality measures QM and scores Sc. Additionally, the linear correlation coefficient between the DR variable and the value of the quality measure gives an indication of the ability of the quality measure to predict errors.

2.2.2. *Not assuming linearity of relationships with quality measures*

In real-world data, the relationship between quality measures QM and scores Sc is not generally linear. This is also observed in [9] for fingerprints. Therefore, we resort to a more sophisticated measure of dependence between these two random variables: the mutual information between score and quality measure $I(Sc;QM)$.

For ease of use in computations and easier interpretability of the measure, we propose to make use a normalised variant of the mutual information, defined by [10]:

$$\bar{I}(Sc;QM) \triangleq \frac{I(Sc;QM)}{\sqrt{H(Sc)H(QM)}}, \tag{1}$$

where $H(Sc)$ and $H(QM)$ are the marginal entropies of scores and quality measures.

2.3. Not assuming homoscedasticity of scores

In practice, it is often found that the variance of scores is largely dependent upon the class. This is explained further in Section 2.4. We amend our basic performance measures to account for this fact.

2.3.1. *Assuming linearity of relationships with quality measures*

The partial correlation coefficient [11] is a modification of the classical correlation coefficient in order to compute the correlation between two random variables given knowledge of the state of another random variable.

The (first-order) partial correlation coefficient is defined as:

$$\rho_{xy\cdot z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}, \tag{2}$$

where the notation $\cdot z$ can be interpreted as "for a subsample where random variable Z has value z". The Z variable is called the control or conditioning variable

To evaluate quality measures, we define two partial correlation coefficients:

$$\rho_{Sc|\Omega} = \rho_{Sc\,QM\,\cdot\Omega} \tag{3}$$
$$\rho_{DR|\Omega} = \rho_{DR\,QM\,\cdot\Omega}, \tag{4}$$

where $\Omega = \{\omega_0, \omega_1\}$ is the class variable representing either clients ω_1 or impostors ω_0.

2.3.2. *Not assuming linearity of relationships with quality measures*

If the linearity assumption is not deemed to hold, as is often the case in real-world data, the partial correlation coefficients should be replaced by a (normalised) conditional mutual information measure obtained on the joint densities of interest, either (Sc, QM) or (DR, QM), defined as:

$$I_{Sc|\Omega} = I(Sc;QM|\Omega) \tag{5}$$
$$I_{DR|\Omega} = I(DR;QM|\Omega), \tag{6}$$

where $\Omega = \{\omega_0, \omega_1\}$ is the class variable representing either clients ω_1 or impostors ω_0.

The contional mutual information $I(X;Y|Z)$ can be interpreted as the mutual information between X and Y, with the effects of the conditioning variable Z removed. We propose a normalised version given by:

$$\bar{I}(X;Y|Z) \triangleq \frac{I(X;Y)}{\sqrt{H(X|Z)H(Y|Z)}}. \tag{7}$$

2.4. The need for class-conditional evaluation of quality measures

While noise can be assumed to have an equivalent effect on impostor and client *likelihoods*, we remark that the relationship between quality measures and *scores* is different for clients and impostors. An example is shown in Figure 1, where a quality measure related to the signal-to-noise ratio is plotted against score distributions. It clearly appears that client scores are more correlated with the quality measures than impostor scores.

In addition to our own experiments, evidence to support this claim is found in numerous publications on speaker recognition. For instance, in [12, Figure 5] it is apparent that the client score distribution is much more affected by mismatched transmission channels than the impostor score distribution. In [13, Figure 1], the addition of artificial Gaussian noise on the speech modality affects the client distribution much more than the impostor distribution. In [14, Figures 2-3], the client distribution is again more perturbed than the impostor distribution when tested with different handsets.

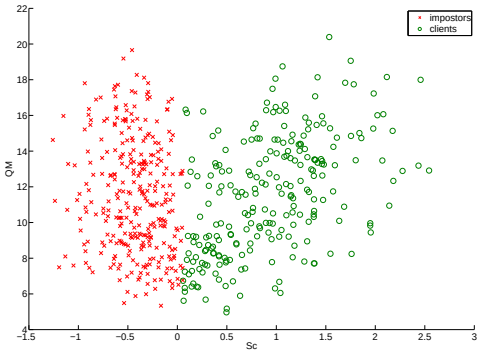

Figure 1: Scatterplot of scores and a SNR-related quality measure showing different correlations depending on class. Crosses indicate impostors and circles indicate clients

This effect is due to the fact that noise manifesting itself as a linear shift in the likelihood domain can affect clients and impostor score distributions differently, because of the logarithmic transform and the world model normalisation used to obtain a log-likelihood ratio.

2.5. A feature selection perspective

An important point is that the ultimate evaluation for a quality measure is to apply it to a biometric verification task dataset and see if it leads to improvements in terms of final error rate or rejection rate. While a quality measure may seem to poorly separate the error-conditional distributions, there may still exist a classifier which can make use of the quality data.

This is analogous to the situation in feature selection: filter methods (functions indicative of the ultimate performance) are generally found to provide inferior results to wrapper methods

[15], where the measure of performance is the use of a feature with the classifier itself.

3. Signal quality measures in speaker verification

Signal quality measures aim to account for degradation in signal quality, which can come from different sources. We can use both time-domain techniques and spectral-domain techniques to obtain a quantity correlated with the amount of noise in the signal.

While much research in speech processing has in the past concentrated on how humans *perceive* speech quality [16], in the application of speaker recognition we are not necessarily interested in trying to emulate human opinion (as represented e.g. by a Mean Opinion Score). Rather, we are looking to design measures corresponding to real-world factors which bear on recognition performance. These two approaches may not yield similar measures.

3.1. Quality measures based on speech segmentation in the time domain

Voice activity detection (VAD), also called speech/pause segmentation, can be used to obtain an estimate of the signal-to-noise ratio. This is done by assuming the average energy in pauses represents the noise energy, and the energy in speech represents the signal energy. In [6], we proposed two algorithms to obtain the segmentation, one based on the energy of the signal, the other based on the spectral entropy. The SNR estimated then yields two quality measures, respectively QM_{VAD_E} and QM_{VAD_H}.

3.2. Quality measures based on higher-order statistics

Since clean speech has a very distinctive distribution (sharp peak at sample value 0 - a large amount of a speech signal is actually silence), we can exploit this knowledge to infer when the signal is noisy. The additive noise we are concerned about has energy (if it does not then it does not impair the speech signal), which means it will contribute to modifying the time-domain distribution of amplitudes.

Higher order statistics can be used to summarise the shape of unimodal distributions in a meaningful way. The skewness (or Fisher skewness) measures the asymetry of a distribution with respect to its mode. Any symetrical distribution (such as Laplace, Gaussian, or uniform) has a skewness of 0. Negative skewness indicates that the distribution has a longer tail on the left of the mode, while positive skewness indicates the opposite.

$$QM_{skew} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{s_t - \mu_s}{\sigma_s}\right)^3, \tag{8}$$

Kurtosis (or Fisher kurtosis), defined in Eq. (9), corresponds to the "peakiness" of the distribution. By definition, a Gaussian distribution has a kurtosis of 3[1]. A leptokurtic (or supergaussian) distribution has a kurtosis higher than 3 and is "peakier", while a platykurtic (or subgaussian) distribution has a kurtosis lower than 3 and is "flatter", that is its probability density is spread over a larger dynamic input range.

[1]Or 0, as some definitions of kurtosis subtract 3 to have kurtosis of 0 for the normal distribution

$$QM_{kurt} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{s_t - \mu_s}{\sigma_s}\right)^4 \tag{9}$$

Unfortunately, kurtosis estimation is very sensitive to outliers. We therefore introduce a third related measure, called the centre bin measure, to approximate kurtosis and estimate the peakiness of the distribution. First, the signal sample amplitudes are binned in 100 equally-spaced bins, then the measure is defined as the ratio of the number of samples in the bin containing the most samples to the total number of samples in the other bins.

$$QM_{bin} = \frac{N_{max}(s)}{\left(\sum_B N_b(s)\right) - N_{max}(s)}, \tag{10}$$

where $N_b(s)$ represents the number of samples in bin b, and $N_{max}(s)$ represents the number of samples in the bin that contains the most samples.

4. Experiments and results

4.1. Systems and databases

The first database is the speech part of XM2VTS , which contains 295 users. The protocol used is the Lausanne protocol, configuration 1. Where applicable, the results are reported by training the models on the evaluation set and testing them on the testing set. We also use a noisy version of XM2VTS, which is generated by adding babble-type noise in SNRs uniformly distributed between 0 and 20 dB.

The second database is the speech part of the BANCA database [17], which contains 2x26 users. The protocole followed is the P protocol. Where applicable, the results are reported by taking an average of measures when first training the fusion model on G1 and testing on G2, then training on G2 and testing on G1.

Additionally, to evaluate the performance of speech segmentation, on which the QM_{VAD} family of quality measures is based, we use the CUAVE audio-visual database [18]. This is a labelled database containing 36 individual users, both male and female, each providing utterances of separated digits for about 2 minutes.

The speaker verification system used for BANCA is based on the Alize toolkit [19]. The Alize speech/pause detector is run to remove silence portions of the input speech signal before feature extraction. Features used are 12 MFCCs with delta and acceleration coefficients, and cepstral mean normalisation. A world model is trained from the pooled clean training data of all clients, using 200 diagonal covariance-matrix Gaussian components. Each client's model is then adapted (means only) with their own recordings using MAP adaptation.

On XM2VTS, we use the 200-Gaussian components GMM classifier from [20], which uses 16 spectral subband centroid features.

4.2. Quality measures based on segmentation in the time domain

4.2.1. Performance of speech segmentation

Since the SNR estimate depends on the speech/pause segmentation, we evaluated the performance of this VAD on the "individuals" set of the CUAVE database. The performance is computed in terms of four quantities [21]: *front-end clipping* (FEC), indicating speech missclassified as noise due to the transition

from noise to speech. *Mid-speech clipping* (MSC) indicates speech misclassified as noise during a speech period. Noise classified as speech when the signal transitions from speech to noise is denoted $OVER$. Finally, noise that is classified as speech during a noise period is denoted NDS. We simplify the evaluation of performance by reporting 3 joint quantities: noise classified as speech ($NAS = OVER + NDS$), speech classified as noise ($SAN = FEC + MSC$), and total error rate R which is the number of signal samples missclassified, no matter whether they were speech or noise. These three quantities are evaluated for each file in the CUAVE database (36 files) and the average is presented in Table 1. It should be noted that the majority of errors are made on three particular files (subjects), and that the files have a high signal-to-noise ratio. Therefore, the VAD will be less accurate on noisy data. This confirms that it could prove useful to combine quality measures derived from this speech/pause segmentation with other quality measures, especially if they are robust to noise (see Section 4.2.2).

NAS_μ [%]	SAN_μ [%]	R_μ [%]
13.03	11.45	12.47

Table 1: percentage of noise samples classified as speech (NAS_μ), percentage of speech samples classified as noise (SAN_μ), and total classification error (R_μ). All results are averaged over the utterances in the individuals set of the CUAVE database.

4.2.2. Performance of SNR estimation

To evaluate the correlation of the energy-based quality measure QM_{VAD_E} with a known signal-to-noise ratio, we run the energy-based VAD algorithm against the noisy version of XM2VTS, thus producing a set (real SNR, quality measure) for each utterance. The results are shown in Fig. 2. Here it can be seen that the energy-based measure is highly correlated ($\rho = 0.82$) with the real signal-to-noise ratio. Thus, it can be expected to be a good indicator of babble-type additive noise.

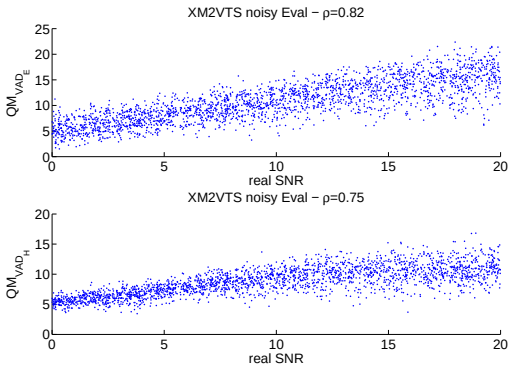

Figure 2: Correlation between the energy-based QM_{VAD_E} signal quality measure and the entropy-based signal quality measure QM_{VAD_H} and real signal-to-noise ratio on a noisy version of the evaluation subset of XM2VTS. Each data point corresponds to an utterance.

Secondly, we run the entropy-based VAD algorithm against the same noisy database to extract QM_{VAD_H}. The results are shown in Fig. 2. The entropy-based measure is also highly correlated ($\rho = 0.75$) with the real signal-to-noise ratio. The superior performance of this estimator in very noisy conditions (SNR=5 dB or below) with respect to the energy-based quality estimator is made clear from this figure, where it can be seen that the spread of estimates for this SNR range is much lower than that of the energy-based quality estimator[2].

4.2.3. Numerical evaluation

As an example of visual inspection, the distribution of the entropy-based quality measure on BANCA is shown in Fig. 3. Here, in general, and according to intuition, higher values of SNR mean higher signal quality and fewer errors. More precise assessment can be obtained by using the numerical performance indicators described in Section 2.3.

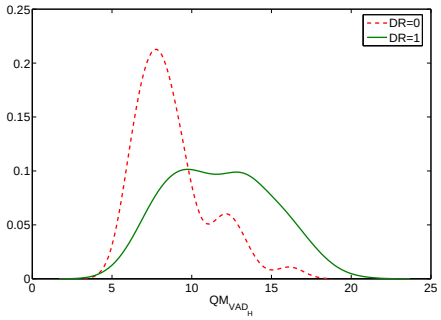

Figure 3: Distributions of entropy-based quality measure QM_{VAD_H} for correct (DR=1) and erroneous (DR=0) classifier decisions on BANCA G1 data.

Table 2 gives results for BANCA, while Table 3 shows results on XM2VTS. The class-conditionality of the relationship between scores and quality measures (pointed out in Section 2.4) is clearly visible from these results: both partial correlation coefficients and conditional mutual information are much larger for clients than for impostors. The second result to point out is that the relationship between scores and quality measures seems stronger than between the reliability indicator DR and quality measures. This can be explained by the fact that the (DR, QM) joint feature space is less informative than the (DR, Sc, QM) feature space, where DR can actually be useful in defining score clusters (erroneous and correct decisions). That is, quality measures on their own, without scores, can not predict errors as well as when used together with scores.

The dependence between scores and quality measures is much less pronounced on the XM2VTS database than on the BANCA database. This is a reflection of the good signal quality encountered in this database, and an indication that quality measures may not improve verification results much in this case.

4.3. Higher order statistics measures of quality

4.3.1. Performance of SNR estimation

To evaluate the correlation of the quality measures with the real signal-to-noise ratio, we again use the noisy XM2VTS. The results for kurtosis (Eq.(9)), skewness (Eq.(8)), and the centre bin

[2] numerically, the residuals for a least-square linear fit are much smaller

measure (Eq.(10)) show that the centre-bin measure is highly correlated ($\rho = 0.54$) with the real signal-to-noise ratio and can be expected to be a good indicator of babble-type additive noise. The Kurtosis is less correlated ($\rho = 0.43$), and the skewness gives negligible correlation ($\rho = 0.17$).

However, good correlation with signal-to-noise ratio does not guarantee that we will be able to predict errors, as the models or features may be somewhat robust to this kind of noise. This is a fundamental point: quality measures must be assessed jointly with *specific* classifier output, and it cannot be said that a particular quality measure is the best to use for all classifiers. Also, it is probable that the best quality measure on a particular database is not the same for other databases, where the noise characteristics may be very different.

4.3.2. Numerical evaluation

We obtain numerical values of the performance of these quality measures; the results are shown in Table 2 for BANCA, and in Table 3 for XM2VTS.

For the BANCA data, it seems that skewness is the most promising of the measures based on higher-order statistics, but that these measures have a weaker dependency relationship with the classifier output than the VAD-based quality measures. However, an advantage is that the distributions of higher-order statistics can be well approximated by a low-order mixture of Gaussians.

For the XM2VTS data, quality measures based on higher-order statistics exhibit higher dependency with classifier output than do VAD-based measures. This is an indication that these quality measures might be favoured for this database.

4.4. Using quality measure in single-classifier speaker verification systems

As a brief example of using quality measures in a speaker verification system, we use a scheme by which a meta-classifier is trained on the score output of the base classifier (see Section 4.1), and this second-level feature space is augmented with a quality measure.

The meta-classifier can be based on generative probabilistic models (as in [6]), on discriminative approaches, or any other classifier. In the current case we use boosting on a C4.5 tree.

The 10-fold cross-valudation results in Table 4 show that all quality measure allow to improve upon the base classifier in terms of Half-Total Error Rate (HTER), even if only slightly. For higher-order statistics, the skewness and center bin measures perform best, as hinted by the generally higher correlations observed in Table 2.

Quality measure	$\Delta_{HTER}[\%]$
QM_{VAD_E}	27.7
QM_{VAD_H}	34.2
QM_{kurt}	11.0
QM_{skew}	34.2
QM_{bin}	26.0

Table 4: Improvement in HTER by using different quality measures on BANCA G2 data. The baseline classifier yields 8.4% HTER on this dataset.

5. Discussion

The results of these experiments point out two important facts. First, that quality measures indeed must be evaluated by taking class-conditionality into account. If not, it might spuriously seem that scores and quality measures are independent. This seems to hold for both well-controlled (XM2VTS) and non-constrained acquisition environments (BANCA). Second, that quality measures really must be evaluated in a classifier and database-dependent fashion. The best quality measure for one classifier might yield very different results on another.

Furthermore, while the evaluation metrics proposed in Section 2 give approximate figures for estimating the usefulness of a quality measure, the ultimate improvement that can be obtained by the use of quality measure depends on the final use of the quality measures: it is not trivial to relate analytically these evaluation metrics to an improved rejection-error tradeoff, or to the final gain yielded by the introduction of quality measures in multiple-classifier fusion.

All proposed quality measures can be used to obtain additional information about the classification problem at hand. It is likely that their combination would bring improvements [6] in terms of error rate.

Depending on the other robustness techniques applied (for instance at the pre-processing stage), the evaluation metrics for the quality measure proposed could yield very different results. Again, the mapping between input signal and output score is complicated analytically by the non-linear transforms intervening in the processing chain, and data-dependent approaches to evaluation of quality measures such as proposed in this paper can be useful additions to the practioner's toolbox.

6. Acknowledgements

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in [22].

7. References

[1] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.

[2] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, June 2006, pp. 1–6.

[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[4] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42–54, Jan. 2000.

[5] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin, "Similarity normalization method based on world model and a posteriori probability for speaker verification," in *Proc. 6th European Conf. on Speech Communication and Technology (EUROSPEECH)*, Budapest, Hungary, Sept. 1999.

measure	$\rho_{Sc\|\omega_0}$	$\rho_{Sc\|\omega_1}$	$I_{Sc\|\omega_0}$	$I_{Sc\|\omega_1}$	$\rho_{DR\|\omega_0}$	$\rho_{DR\|\omega_1}$	$I_{DR\|\omega_0}$	$I_{DR\|\omega_1}$
QM_{VAD_E}	-0.18	0.49	0.04	0.15	0.28	0.11	0.08	0.04
QM_{VAD_H}	-0.17	0.48	0.04	0.12	0.27	0.09	0.07	0.04
QM_{skew}	-0.06	0.31	0.03	0.09	0.05	0.04	0.06	0.01
QM_{kurt}	-0.04	0.16	0.03	0.08	-0.02	0.07	0.09	0.01
QM_{bin}	-0.11	0.27	0.04	0.06	0.10	0.06	0.06	0.02

Table 2: Average performance of signal quality measures on the BANCA dataset.

measure	$\rho_{Sc\|\omega_0}$	$\rho_{Sc\|\omega_1}$	$I_{Sc\|\omega_0}$	$I_{Sc\|\omega_1}$	$\rho_{DR\|\omega_0}$	$\rho_{DR\|\omega_1}$	$I_{DR\|\omega_0}$	$I_{DR\|\omega_1}$
QM_{VAD_E}	0.09	0.34	0.03	0.06	-0.05	0.15	0.01	0.03
QM_{VAD_H}	0.02	0.27	0.03	0.05	-0.03	0.11	0.01	0.02
QM_{skew}	0.08	0.52	0.03	0.14	-0.03	0.14	0.01	0.04
QM_{kurt}	0.07	0.43	0.03	0.09	-0.03	0.11	0.01	0.02
QM_{bin}	0.09	0.41	0.02	0.08	-0.05	0.10	0.01	0.02

Table 3: Performance of signal quality measures on the XM2VTS evaluation dataset.

[6] Jonas Richiardi, Plamen Prodanov, and Andrzej Drygajlo, "Speaker verification with confidence and reliability measures," in *Proc. 2006 IEEE International Conference on Speech, Acoustics and Signal Processing*, Toulouse, France, May 2006.

[7] Ji Mingy, Timothy J. Hazenz, and James R. Glassz, "A comparative study of methods for handheld speaker verification in realistic noisy conditions," in *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

[8] D. Garcia-Romero, J. Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez, and Javier Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," in *Proc. ODYSSEY 2004 - the Speaker and Language Recognition Workshop*, Toldeo, Spain, May-June 2004, pp. 105–110.

[9] Patrick Grother and Elham Tabassi, "Performance of biometric quality measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

[10] Alexander Strehl and Joydeep Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. of Machine Learning Research*, vol. 3, pp. 619–620, 2002.

[11] David J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC press, 2004.

[12] Johan Koolwaaij and Lou Boves, "On decision making in forensic casework," *The International Journal of Speech, Language and the Law: Forensic Linguistics*, vol. 6, no. 2, pp. 242–164, 1999.

[13] Sonia Garcia-Salicetti, Mohamed Anouar Mellakh, Lorene Allano, and Bernadette Dorizzi, "Multimodal biometric score fusion: the mean rule vs. support vector classifiers," in *Proc. 13th European Signal Processing Conference (EUSIPCO)*, 2005.

[14] L.P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1997, vol. 2, pp. 1071–1074.

[15] George H. John, Ron Kohavi, and Karl Pfleger, "Irrelevant features and the subset selection problem," in *Proc. Int. Conf. on Machine Learning (ICML)*, 1994, pp. 121–129.

[16] A.W. Rix, J.G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality - technology and applications," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1890–1901, November 2006.

[17] Enrique Bailly-Bailliére, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belen Ruiz, and (Jean-Philippe) Thiran, "The BANCA database and evaluation protocol," in *Proc. 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, J. Kittler and M.S. Nixon, Eds., 2003, vol. LNCS 2688, pp. 625–638.

[18] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1189–201, Nov. 2002.

[19] Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005, pp. 737–740.

[20] Norman Poh and Samy Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, no. 2, pp. 223–233, Feb. 2006.

[21] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1989, vol. 1, pp. 369–372.

[22] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 1999, pp. 72–77.