ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **17**, Number 2, 2014, 134–149

A Quality-Aware Forensic Speaker Recognition System

Gheorghe POP[1], Dragoş DRĂGHICESCU[2], Dragoş BURILEANU[2]

[1] National Institute for Forensic Expertise, Bucharest, Romania
[2] University "Politehnica" of Bucharest, Faculty of Electronics, Telecommunications and Information Technology

E-mail: gheorghe.pop@inec.ro

Abstract. The performance of a speaker recognition systems based on Gaussian mixture models is often impaired both by the low quality and by short duration of test speech samples. A best material selection criterion is described in this paper, especially suitable for forensic automatic speaker recognition systems, where even enrollment speech quality might be important in some cases. The material selection is performed by checking well known short-time measures of input speech that carry quality information, such as linear cepstral peak, spectral autocorrelation peak to valley ratio, and windowed autocorrelation lag energy. Our tests show that the proposed approach outperforms reported speaker recognition solutions that consider quality of the input speech, at least in co-channel speech conditions.

Key-words: Forensic automatic speaker recognition, Gaussian mixture model, quality measure, linear cepstral peak.

1. Introduction

A forensic automatic speaker recognition (FASR) system can be thought as quality-aware if assistance with data about the quality of the input speech is implemented, at least in front end and score calculation stages. Traditionally, best input material selection was addressed by detection of voice activity (VAD) in questioned utterances, although quality of speech has always been central to the success of speaker recognition solutions.

In spite of the huge steps both digital signal processing and automatic speaker recognition systems have taken in the latest decades, as well as their increasing number of applications, they are still challenged by the ability of humans to extract robust information from speech.

While in telecommunications quality of speech signal is understood in relation to the average listener, forensic examiners define it for the recognition task at hand. Meanwhile, in biometrics, quality measures are defined as *information that helps assessing the probability that a biometric verification decision is correct*.

Over time, speaker recognition was performed in several scenarios, although most of them are different flavors of *speaker identification*, and *speaker verification* (or *authentication*). Identification consists of establishing the identity of a person from a closed or open set of suspected persons, while verification is the equivalent for just one suspect. A few varieties of speaker recognition, such as *detection*, *authentication*, and *diarization*, have different names that come from their applications. *Speaker detection* aims to discover the presence of one or more listed persons into a supervised space, while *speaker authentication* means to attach the credentials given in relation to a protected resource, to a person whose utterances were checked to correspond to a claimed identity. *Speaker diarization*, instead does no mandatory association of speakers to known identities, but creates a diary of the contents of a speech recording in a *who spoke what* manner. Speaker diarization may be seen as doing open set speaker identification of each speaker in input audio, with a supplemental task to discriminate each unknown speaker from all others, and assign them a different, automatically generated, label.

From a forensic point of view, all the scenarios described here are now acceptable in the police inquiries, while in court they are definitely obsolete, mostly because their decisions were categorical. A simple reason stands behind this statement, which owes historical roots to a warning that the National Academy of Sciences in United States of America has issued in 1979, against the use in court of analysis methods that were not scientifically validated. Speaker's uniqueness and the relevance of his uniqueness were at stakes, and led to a new paradigm in forensic speaker recognition.

An important difference from biometric to forensic speaker recognition comes from the fact that utterances in biometrics may sometimes be requested twice or more, while in forensics there is no such thing as collecting a new questioned utterance when quality of given speech is not good enough.

Therefore, the best use of the acquired speech signal in forensic scenarios relies on giving each short-time frame of the input signal the appropriate weight, which relates to the true representation of the target speech under the given conditions of signal-to-noise ratios (SNRs), channel artifacts, modality of collection, etc. Mismatch is very likely to occur between conditions in which audio materials were acquired, given that test recording might originate from lawful intercept of speech communications while the enrollment speech recordings come from interview sessions. Low speech quality and short duration of available pure speech thus pose problems to speaker recognition systems, similar in difficulty to channel mismatch and within-speaker variability.

2. Overview on speech quality research

After the idea of speaker recognition based on Gaussian mixture models (GMMs) was introduced in the 90 s [14], many of its variations have been developed to address factors involved in the performance degradation of automatic speaker recognition systems. One such factor has always been the input speech quality. Accuracy and reliability of speaker recognition systems are shown by indicators like the equal error rate (EER), the detection error tradeoff (DET) curve, the minimum value of the detection cost function (minDCF), and a few others.

GMMs were first used in conjunction with cepstral coefficients, both in linear and mel-warped frequency scales. These are generative models, thus they also account for some values of the feature vectors that were not seen in their training data.

Mainly due to the lack of information relevant to the speaker identity, a decrease in recognition performance occurs in FASR systems for low quality speech or short utterances. More and more speech samples from low quality questioned files are excluded in these cases, or just labeled as inappropriate for the specific task. It is now common practice to run recognition processes only after there is enough pure speech available for both enrollment and test phases. Enrollment phase is usually controllable since enough high quality material is available in most cases. An additional problem in forensics comes from the possible hostility of the enrolled speaker, who might make efforts to spoil the acquired material.

Extensive studies were reported in [5] and [10] on the discrimination performance of FASR systems with short duration input recordings, using utterance length as a quality measure function (QMF). In [7], universal background model (UBM) misalignment was used as a quality measure, as it mostly indicates specificity of the measured feature vector.

Other unimodal biometric QMFs may be defined, such as a distance of their argument to a user-independent decision threshold, that are absolute or relative, modality dependent (signal-domain) or independent (at the cost of providing lower recognition accuracy), score-based or user model-based. Combining these quality measures with score-based quality measures could simplify the subsequent classification or regression task. The authors of [15] have shown that to obtain better modeling of error conditions, several quality measures should be combined. A score-based quality measure alone may not lead to high accuracy in recognizing error conditions, but combining it with a distance to a decision threshold yields much better results. Likewise, adding an entropy-based quality measure for speech helps compensate the drawbacks of energy-based quality measures in high noise situations. To the same end, quality measures may be defined that are themselves arithmetic aggregates of other quality measures, each one accounting for a different aspect of the signal.

The output of a model that comprises quality information can be used afterwards for either automatic matching algorithm or human examination [2].

In [6] a probability-theoretic framework was described for defining quality measure functions as the degree of goodness for each short-time frame of speech, given a certain criterion. It is a sensible idea that best material selection could be implemented using QMFs that they take real values from 0 to 1. While 1 is intuitively associated with the

full compliance to the goodness criteria, 0 would indicate failure to reach the minimum goodness (henceforth called *acceptance criteria*). This approach would hold a bit of confusion. Probabilities are always positives equal to 1 or less, so combining various such quality measures by multiplication or raising to power would also give positive values equal to 1 or less. The results of such calculations would give results closer to 1, thus making the speaker recognition features count more, in fact, when quality measures are lower, which should not be opposite to their role.

An evaluation of quality measures was proposed in [15] based on their distribution in two classes: DR = 1 (Decision Reliable) and DR = 0. To evaluate the ability of a quality measure to predict errors, a linear correlation coefficient has to be computed between quality measures and recognition scores (under linearity assumption). It follows that, after discarding from the test signal all portions with quality under acceptance criteria and adjusting the weight of remaining material to better promote those with higher quality, the pure speech equivalent (PSE) duration of usable speech should be updated accordingly.

In [11], pure voice duration and SNR were used as quality measure functions, and their use for calibration of two state-of-the-art systems, i-vector and probabilistic linear discriminant analysis (PLDA) was described for short test utterances. These QMFs are though not designed to handle the condition where duration goes to zero. Recognition systems get no speaker information, hence both hypotheses are equally likely in this extreme case.

The introduction of speech quality information in speaker recognition was considered in literature at various levels and in various fusion schemes. Two sound theoretical frameworks for combining several machine experts, each using only one quality-related parameter, were described in [1] and [8] for speaker authentication. Recent studies have also found that reliability of the results deteriorates when combining multiple speaker recognition systems using logistic-regression [4].

From a forensic point of view, quality of speech should be defined in order to provide at least:

1. a decision whether enough information relevant to the target speaker identity exists in the questioned recording. If a decision was made to pursue the assessment of speaker identity, then quality definition would provide two more requirements:

2. a criterion for optimal selection of input speech material, and

3. a quality-aware speaker recognition evidence to be derived and output from FASR systems.

If several quality-related features are available for the same short-time interval, they would form a vector which may be called a *q-vector*. A block diagram of a q-vector-based case pre-assessment system is shown in Fig. 1, which could be used to generate the decision (1), as detailed in [13].

Solutions for the other two criteria may be implemented into quality-aware FASR. Details of such systems based on Gaussian mixture models with universal background

model (GMM-UBM) are given in literature; they deal with some known quality-related factors by either adaptation of the background speaker model or by establishing priors used in the Bayesian inference stage [16].

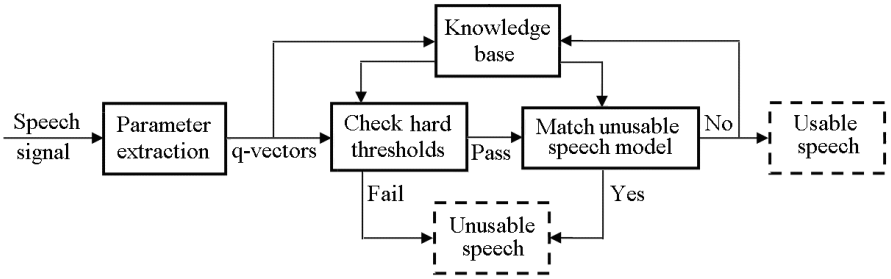

Fig. 1. Block diagram of a case pre-assessment stage in FASR.

Assessing quality of recorded speech by global quality descriptors can raise important issues in forensics, as the region in the recording that holds the target speech can have very different quality measures, which may render the global quality cues largely irrelevant. A *model of unacceptable speech* could be trained, for use with such a pre-assessment stage. An extended set of local quality measures can be used to statistically model the *total quality*. For example, fundamental frequency (F_0) deviations, SNR and ITU-T P.563 objective quality scores were used in [6] as goodness criteria in a multilevel speaker recognition system (this solution will be analyzed in Section 6).

3. The proposed approach to front end quality features

Most of the time, the concept of *usable speech* has been specifically defined in literature to work at the pre-processing stage, within the speaker recognition system of choice. In our pre-processing approach no enhancement is allowed on the input speech, other than the compensation of recording conditions mismatch, such as channel effects and acoustic reverberation. One of the most important problems in forensic speaker recognition was perhaps the duration of available speech, which has fueled a strong competition between speaker recognition systems towards the setting of a minimum duration limit, at least for questioned utterances.

Quality measures such as SNR, reverberation time (T_{60}), and pure speech duration were long ago introduced as acceptance criteria to the input signal. As, in many cases, the global quality descriptor approach is not such a good idea, the quality measures and speaker recognition features are to be extracted for each short-time analysis frames of about 20 ms, with an overlap of 10 ms, and used in order to comply to the three requirements stated in the previous section for such systems. The front end process is equivalent to a case pre-assessment stage and works just as described in Fig. 1. The output of the pre-assessment process is the usable speech, formatted as a feature vector series, and the knowledge base, which mainly includes the q-vector series.

Quality-related speech parameters may have their bad influence either at lower values, average values or high values, while in other ranges may lack effects. On the other hand, there could be a sharp transition or a smooth one, between the extreme cases; however, useful thresholds can be set experimentally, for any given quality feature.

Features of interest for quality description must be highly discriminative while being very robust to different conditions. Thus we selected short-time measures related to speech quality that could be useful in detection of unusable portions of speech, such as co-channel speech and high levels of noise.

We chose the following individual quality measures:

1. *inverse linear cepstral peak* (ILCP),
2. *log windowed autocorrelation lag energy* (logWALE) [9], and
3. *modified spectral autocorrelation peak to valley ratio* (MSAPVR) [3].

To reduce correlation between the selected measures, we left aside excess kurtosis, which is good on noisy signal detection, but nothing was lost, as log-kurtosis is highly correlated to logWALE, which we preferred. In Fig. 2, an illustration of this situation is shown, for a randomly selected 10 seconds speech file.

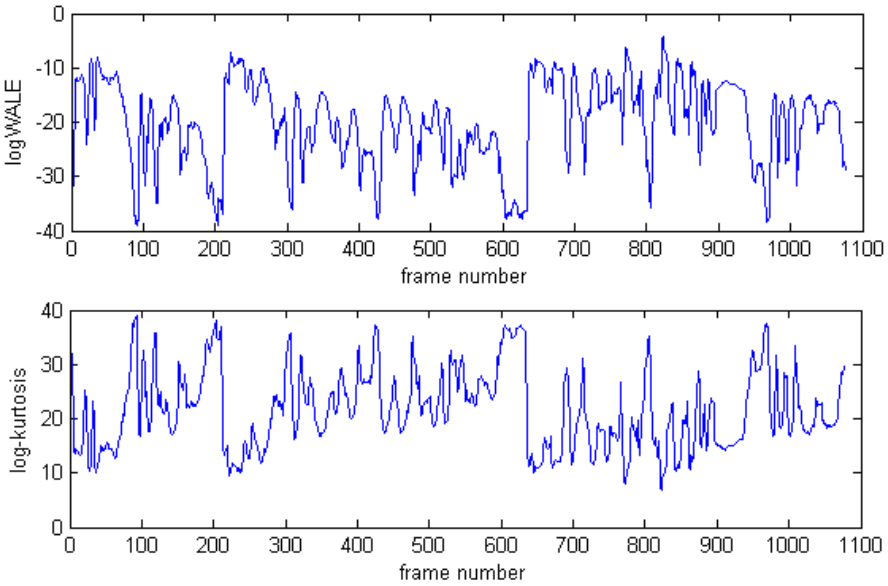

Fig. 2. High correlation of logWALE and log-kurtosis.

ILCP is the reciprocal of another known measure, linear cepstral peak (LCP), used for pitch estimation and voice activity detection, and as an indicator related to the noise floor of the signal.

LogWALE is a quality measure that attempts to condense the harmonic structure of voiced speech into a single coefficient that is relatively easy to compute. It is robust for noise and voiced speech distributions, close-talk and far-talk, and produces lower rate of usable speech false rejects. It is also gain independent, as it uses normalized autocorrelation.

The MSAPVR measure is a modification of the spectral autocorrelation peak to valley ratio (SAPVR), defined in [9], which targets the LPC residual of the signal instead of the signal itself, in order to emphasize the existence of a speech-specific spectral structure, which almost vanishes in co-channel speech conditions.

In our approach, the domain of each selected quality feature does not natively span between 0 and 1, so a function is needed to map them to the set of positive real numbers lower than 1. Such functions would be authentic QMFs, and multiple quality functions are easier to combine, in this numeric setup. Based on the selected underlying measures, the q-vector components will take the values of QMFs for each short-time frame. As a fused quality indicator, we used the frame wise product of all available QMFs, in both the front end and score computation stages.

At the front end stage, the ILCP quality feature time series was computed as

$$\text{ILCP} = \{\max_{n}(\text{dct}(\log(|\text{FFT}(y_t)|^2)))\}^{-1};\ t = 1, 2, \ldots, T, \tag{1}$$

where y_t is the input signal over the short-time frame t. Low values of ILCP show noisy signals, while high values indicate cepstrum structure.

The logWALE time series was obtained as

$$\text{logWALE} = \{\log(\max_{k}(\sum_{i=k}^{k+W-1} |\text{accor}_t(i)|^2))\};\ t = 1,2,\ldots,T, \tag{2}$$

where W is the width of the lag window in which the maximum autocorrelation is searched. In [9], a value of $W = 15$ is recommended, which we considered acceptable.

MSAVPR was first introduced as a means to detect usable speech even in co-channel speech conditions. When using it as a quality measure, co-channel speech should be excluded for target-to-interferer ratio (TIR) below 20 dB. We adopted the modification proposed in [3], which targets linear prediction (LP) residual. It has the advantage of being highly periodic because the vocal tract information is almost removed from the original signal.

The time series for the MSAPVR measure was calculated as the ratio of the sum of twice the first peak and next four peaks to the first valley:

$$MSAPVR = \left\{ \left(\frac{2 \times P_1 + P_2 + P_3 + P_4 + P_5}{V_1} \right) \right\}, \tag{3}$$

where P_1, P_2, P_3, P_4, and P_5 are the first 5 peaks in the spectrum autocorrelation of the LP residual, and V_1 is the first valley.

By examining the time series (1), (2), and (3), various dynamic ranges were observed. In order to obtain the desired quality measure functions, we clipped their dynamic range at experimental thresholds, then mapped the range to [0; 1]. When a

short-time signal frame has any of the quality measures lower than 1/1,000,000, it is rejected as unusable. The weight of each usable speaker feature vector is set as the product of the defined QMFs, to better promote features extracted from high quality frames, and prevent features from low quality frames from having destructive effects on the speaker recognition outcome.

Compliance to the first requirement discussed in Section 2 asks FASR to also assess the pure speech equivalent (PSE) duration of the accepted speech, from which we could learn the relevance of the questioned speech file for speaker recognition. At this point, we made a presumption that the contribution of each frame to the PSE was proportional to the reciprocal of a fuzed quality-measure.

When quality signals depart from the maximum, the FASR system reacts by reducing the weight of the corresponding speaker recognition feature vector in computing the likelihoods of the test utterance, given both the UBM and the hypothesized GMM speaker model. Therefore any global quality problems should be addressed before speaker recognition, through processes like UBM compensation, then begin adaptation of speaker models with respect to an already compensated background model.

4. Quality-based score computation

The proposed approach allows easy inclusion of new quality measures at any later time, under the condition of:

1. adding a frond-end module, which extracts the quality measure values and a quality measure function which monotonically maps the range of the measure into [0; 1], and
2. adjusting the acceptance criteria and calculations of the recognition scores accordingly.

In the following, we describe the use of QMFs in score computation. We denote the consecutive short-time analysis frames as

$$\mathbf{Y} = \{y_t\}, \quad t = 1, 2, \ ..., T. \tag{4}$$

To allow for the usage of k QMFs, both at the best material selection and score computation stages, we considered the time series of each QMF over the entire timeline of the questioned audio, which is sometimes called a *quality signal*,

$$\mathbf{q}^k = \{q_t^k\}, \quad t = 1, 2, \ ... \ , T; \quad k = 1, 2, \ ... \ , N, \tag{5}$$

given a GMM speaker model, λ, with M d-variate Gaussian components,

$$\lambda = \{w_i, \mu_i, \mathbf{\Sigma}_i\}, \quad i = 1, 2, \ ... \ , M, \tag{6}$$

where$w_i$are the weighting coefficients, μ_i are the mean vectors of the i-th Gaussian component in the mixture, and $\mathbf{\Sigma}_i$ the covariance matrix of the feature vectors along the d dimensions.

The likelihood function of a feature vector (seen as a d-dimensional observation, $\mathbf{o}$), generated by the model λ, is then

$$p\left(\mathbf{o}|\lambda\right) = \sum_{i=1}^{M} w_i p\left(\mathbf{o}|\mu_i, \mathbf{\Sigma}_i\right). \tag{7}$$

Given a speech signal, $\mathbf{Y}$, and a sequence of feature vectors, $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ ... \ , \mathbf{o}_T\}$, which may be considered independent, a time series may be created for quality measure k, $\mathbf{q}^k = \{q_1^k, q_2^k, \cdots, q_T^k\}$, from the values of the quality measure in all short-time frames by a mapping reciprocal to the one we described previously, so $q_t^k \in [0;\ 1]$.

The likelihood of a model λ that incorporates the quality measure k, for the sequence of feature vectors, will be:

$$p\left(\mathbf{O}|\mathbf{q}^k, \lambda\right) = \prod_{t=1}^{T} p\left(\mathbf{o}_t|\lambda\right)^{q_t^k}. \tag{8}$$

A high value of q_t^k makes $p\left(\mathbf{o}_t|\lambda\right)^{q_t^k}$ much closer to 1, which reduces the contribution of $\mathbf{o}_t$ to the global likelihood.

Combining and incorporating quality information from front end processing tends to over-penalize portions of speech signals which have more than one quality measure lower than 1. The inclusion of N quality measures is especially important when their values are low, because models of good and bad quality may be defined in an N-dimensional vector space.

Ideally, model adaptation should be based upon portions of data that also increase the quality of the model. This is why we chose to use the quality signals as powers of the likelihoods. The log-likelihood can then be computed as

$$\log p\left(\mathbf{O}|\mathbf{q}, \lambda\right) = \sum_{t=1}^{T} V\left(\mathbf{q}_t\right) \log p\left(\mathbf{o}_t|\lambda\right), \tag{9}$$

where $V\left(\mathbf{q}_t\right)$is the volume the q-vector stretches upon, at time t, in the quality space. We chose this as a means to normalize the feature vector contributions, which should meet the stochastic requirement:

$$\sum_{t=1}^{T} V\left(\mathbf{q}_t\right) = 1. \tag{10}$$

Reasonable use of input material has to balance reliability and quality. To this end, the contribution of each input frame to the PSE duration would be the frame duration divided to its global quality measure, the reciprocal of the volume $V\left(\mathbf{q}_t\right)$ of the q-vector.

5. Experiments

Our experiments started from a baseline FASR, built around the Microsoft speaker recognition (MSR) identity toolbox [12], which was a low level, text-independent one, based on GMM-UBM. Some of the information needed to implement the previous equations was obtained experimentally.

First, we tested the rejection power of the quality features. For example, in Fig. 3, the ILCP time series, and the log-likelihood ratio (log-LR, or LLR) are shown, in alignment to the signal waveform, for a randomly chosen short speech file.

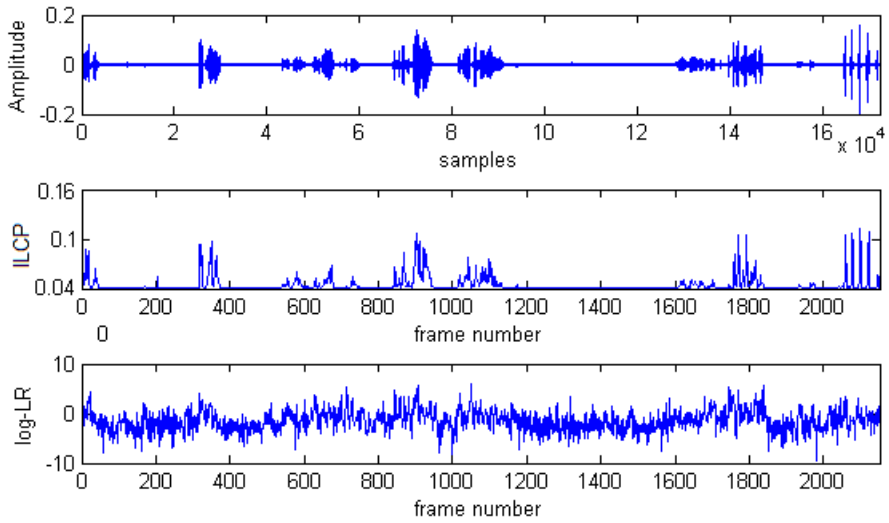

Fig. 3. Plots of the waveform, the corresponding ILCP, and log-LR speaker recognition score time series for a short speech file.

The log-LRs are visibly pushed down in the signal regions that have no speech. It is common sense that lack of quality in input speech should lead to lower reliability of the evidence pertaining to the questioned speaker.

The performance of the baseline FASR system is demonstrated in Fig. 4 for speech recordings from the NIST 2008 SRE database. A number of 250 speech files, recorded with 8 kHz, 16 bits per sample, were randomly selected from the *10 seconds* train section, and a 42 mel-frequency cepstral coefficients (MFCCs), 64 Gaussians UBM was trained from the selected files. From files unseen in the UBM training, 34 speaker models were obtained by maximum a posteriori adaptation from the UBM, and 34 test files were preserved for FASR evaluation. In Fig. 4 the confusion matrix obtained after the system was tested with the 34 speakers and the corresponding detection error tradeoff (DET) plot are shown (the hotter the color in the confusion matrix, the higher the log-likelihood that the speech in the file on the same line was generated by the model on the same column).

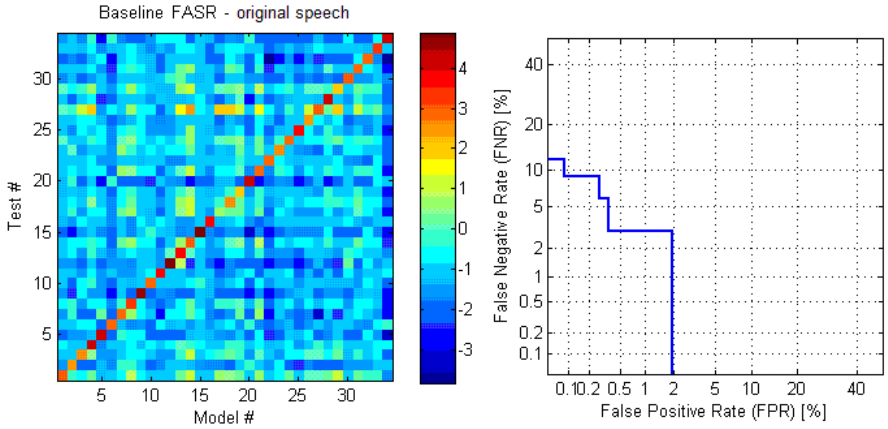

Fig. 4. The confusion matrix and DET plot for the baseline FASR system – original speech.

In a third experiment, ILCP, logWALE, and MSAPVR features were examined for a number of 150 files, each around 8 minutes long, from NIST 2008 SRE speech database. In order to develop intended quality measures, proper thresholds were searched by examining the statistics and correlations of their values with speaker recognition scores.

In Fig. 5, normalized histograms of the three selected quality measures and their fused values are shown for a typical file, where $\mathbf{q}^1$, $\mathbf{q}^2$, $\mathbf{q}^3$, and $\mathrm{V}(\mathbf{q})$ are the three QMFs, and their fused value, respectively.

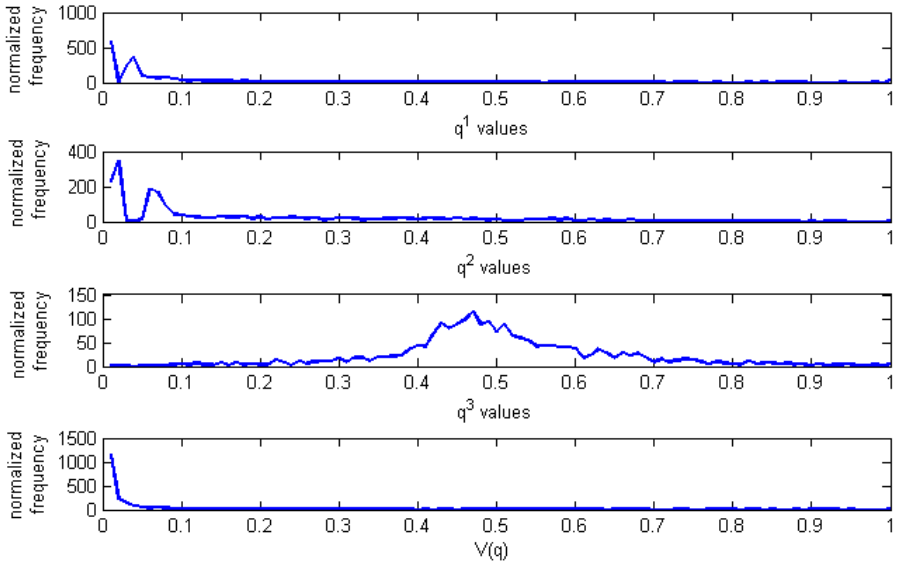

Fig. 5. Normalized histograms of the three QMFs, and their fused value in a randomly chosen file, typical for FASR systems.

Based on the statistic behavior observed in individual quality measures, the needed thresholds were chosen and the quality measure functions resulted as:

$$\mathbf{q}^1 = (10 \times \mathrm{ILCP})^4 , \tag{11}$$

$\mathbf{q}^2 = 1 - 25 \times \log \mathrm{WALE}$, and (12)

$$\mathbf{q}^3 = 1 - \sqrt{\mathrm{MSAPVR}}. \tag{12}$$

We conducted a fourth experiment, in order to demonstrate the impact of quality awareness of the FASR system on the speaker recognition performance. The same 34 test files that were preserved for the baseline FASR system evaluation, were now mixed with babble noise, adjusting the gain of the babble noise file in order to insure a specific target-to-interferer ratio (TIR). The confusion matrix and DET plot obtained using the baseline FASR system are shown in Fig. 6 for the babble-contaminated versions of the evaluation files, with TIR = 15 dB.

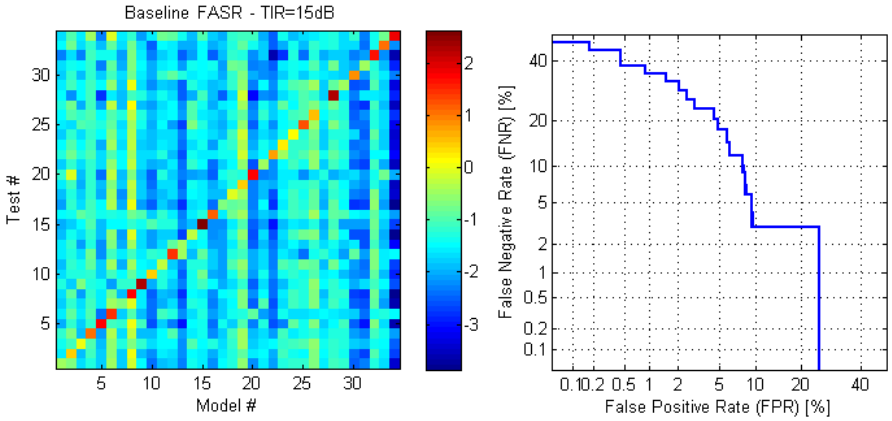

Fig. 6. The confusion matrix and DET plot for baseline FASR, with babble-noise speech – TIR = 15 dB.

Using the same UBM and GMMs, the evolution of the confusion matrices is depicted in Fig. 7 from baseline to quality-aware system, for the 34 evaluation files (original speech).

TIR value lower than 20 dB would normally render unusable all the content of the test files. As Fig. 6 shows, there are obvious cases where speakers known to the FASR system are not recognized from their noisy speech. Speaker feature vectors are altered by the babble, because of its similarity to the target signal, so there would be little to tell about the target speaker. However, there might be short time frames with a single dominant speaker. The introduction of quality measures into the score computation makes it possible for the dominant speaker to be recognized even at TIR values lower than the well-established 20 dB threshold, as Fig. 8 shows.

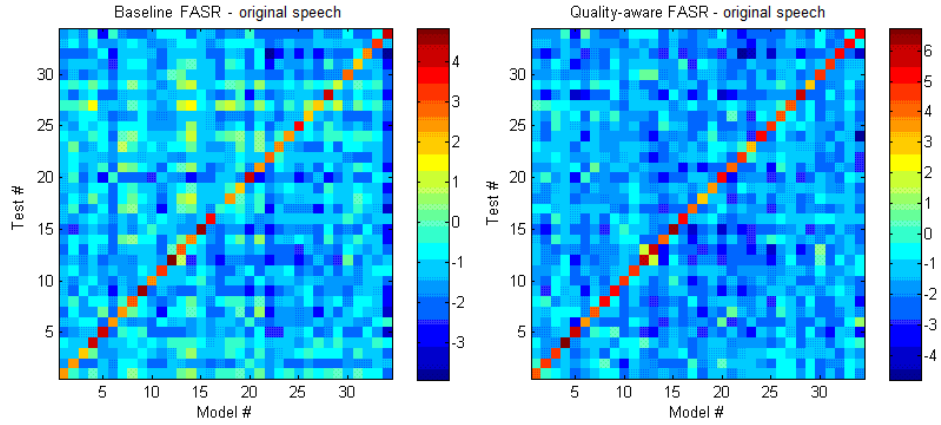

Fig. 7. The evolution of the confusion matrix from baseline to quality-aware system – original speech.

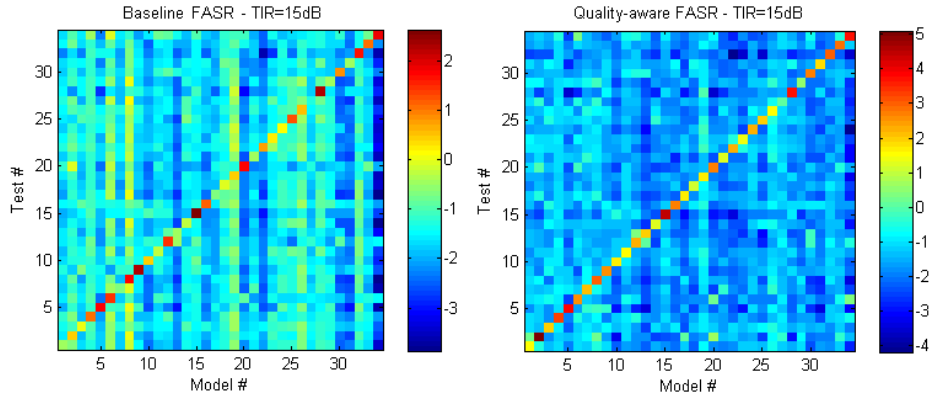

Fig. 8. The evolution of the confusion matrix from baseline to quality-aware system – TIR = 15 dB.

The performance of the described quality-aware FASR system is also illustrated in Fig. 8, for the same evaluation files, after they were contaminated with babble noise, at TIR = 15 dB, by the evolution of the confusion matrix from baseline to quality-aware system.

The PSE duration of the pure speech is computed by summing equivalent durations of each short-time frame. For users of FASR systems, who need to know if a speech file is usable before having it tested against suspect speakers, PSE duration thresholds could be found by building a diagram of the recognition scores as a function of PSE duration, and picking from there the point where the recognition scores stop growing with further increase of duration.

6. Conclusions

In this paper, a quality-aware forensic speaker recognition system was presented, and its performance with both clean and noisy speech was discussed. We compare our system to the one described in [6] which is similar in the short-time scale quality data, and in using score fusion.

Quality measures such as UBM misalignment and F_0 deviations are relative to an artificial data-driven model of the signal quality. The weight of each feature vector in the recognition score, in [6], depends on the context-specific average F_0. By doing so, a slow varying, steady, or robotic voice obtains better scores with such systems than normal voice of the same speaker would do. In a similar manner, a feature vector that fits in the UBM at a greater distance from the mean values of the Gaussians, could be regarded as having a higher quality, even if it is noisy or co-channel. There is in fact an important quality of such feature vectors: they carry more specificity, and based on that there is a lower probability for them to be observed by pure chance.

Our approach outperformed the results reported in [6], as shown in Table 1.

Table 1. Performance of the system compared to the one described in [6]

Performance indicator	**Alternative approach [6]**	**Proposed approach**
Nature of underlying features	3 low level; 2 high level	3 low level
Granularity	F_0 – granular SNR and P.563 – global	Completely granular
Automation	Not complete	Complete
EER improvement from baseline	7.15% for multilevel system	At least 15% for single level system

The system we propose is more automatic, more granular than alternatives in [7], where UBM misalignment is used, and [6], which uses F_0 deviations.

Our FASR system performs by more than an order of LR magnitude better than other approaches, from which only the one in [6] is similar in principle. With co-channel speech, our system was proven efficient even at TIR of 15 dB, which is lower than the generally accepted limit of 20 dB. Besides the F_0 deviations, the similar quality approach in [6] has used quality descriptors such as SNR, and objective quality of the signal, defined in the ITU-T P.563 recommendation. As shown in the paper, the quality-aware FASR system we propose has better accuracy and reliability than both the baseline system and the approach in [6]. However, the latter is the only work that takes a comparable view on speech quality. Higher performance of our system was brought by a better scoring, and the ability to penalize contributions of noisy and co-channel speech.

References

[1] BIGUN E.S., BIGUN J., DUC B., FISCHER S., *Expert Conciliation for Multi-Modal Person Authentication Systems by Bayesian Statistics*, Proceedings of IAPR Interna-

tional Conference on Audio and Video-Based Person Authentication (AVBPA), Springer LNCS, pp. 291–300, 1997.

[2] CAMPBELL W.M., REYNOLDS D.A., CAMPBELL J.P., BRADY K.J., *Estimating and Evaluating Confidence for Forensic Speaker Recognition*, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, vol. **1**, pp. 717–720, 2005.

[3] CHANDRA N., YANTORNO R.E., *Usable Speech Detection Using Spectral Autocorrelation Peak to Valley Ratio Using the LPC Residual*, Proceedings of the 4^{th} IASTED International Conference of Signal and Image Processing (SIP), Kauai, Hawaii, pp. 146–150, 2002.

[4] ENZINGER E., MORRISON G.S., *The Importance of Using Between-Session Test Data in Evaluating the Performance of Forensic Voice-Comparison Systems*, Proceedings of the 14^{th} Australasian International Conference on Speech Science and Technology, Sydney, pp. 137–140, 2012.

[5] FAUVE B., EVANS N., PEARSON N., BONASTRE J.F., MASON J., *Influence of Task Duration in Text-Independent Speaker Verification*, Proceedings of the 8th Annual Conference of the International Speech Communication Association (ISCA) – INTERSPEECH, Antwerp, vol. **7**, pp. 794–797, 2007.

[6] GARCIA-ROMERO D., FIERREZ-AGUILAR J., GONZALEZ-RODRIGUEZ J., ORTEGA-GARCIA J., *Using Quality Measures for Multi-Level Speaker Recognition*, Computer Speech and Language, vol. **20**, no. 2–3, pp. 192–209, 2006.

[7] KELLY F., DRYGAJLO A., HARTE N., *Compensating for Ageing and Quality Variation in Speaker Verification*, Proceedings of the 13^{th} Annual Conference of the International Speech Communication Association (ISCA) – INTERSPEECH, Portland, Oregon, Sept. 9–13, 2012.

[8] KITTLER J., HATEF M., DUIN R., MATAS J., *On Combining Classifiers*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. **20**, no. 3, pp. 226–239, 1998.

[9] KRISTIANSSON T., DELIGNE S., OLSEN P., *Voicing Features for Robust Speech Detection*, Proceedings of the 6^{th} Annual Conference of the International Speech Communication Association (ISCA) – INTERSPEECH, Lisbon, pp. 369–372, 2005.

[10] MANDASARI M.I., MCLAREN M., VAN LEUWEN D., *Evaluation of i-vector Speaker Recognition Systems for Forensic Application*, Proceedings of the 12^{th} Annual Conference of the International Speech Communication Association (ISCA) – INTERSPEECH, Florence, pp. 21–24, 2011.

[11] MANDASARI M.I., SAEIDI R., MCLAREN M., VAN LEUWEN D., *Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions*, IEEE Transactions on Audio, Speech and Language Processing, vol. **21**, no. 11, pp. 2425–2438, 2013.

[12] MICROSOFT, *MSR Identity Toolbox (With Binaries)*, v1.0, `http://research.microsoft.com/en-us/downloads/2476c44a-1f63-4fe0- b805-8c2de395bb2c/`, published 17 Oct. 2013.

[13] POP G., DRĂGHICESCU D., BURILEANU D., *On Forensic Speaker Recognition Case Pre-Assessment*, Proceedings of the 7^{th} Conference on Speech Technology and Human-Computer Dialogue (SpeD), Cluj-Napoca, pp. 169–176, 2013.

[14] REYNOLDS D.A., *Speaker Identification and Verification Using Gaussian Mixture Speaker Models*, Speech Communication, vol. **171**, no. 2, pp. 91–108, 1995.

[15] RICHIARDI J., DRYGAJLO A., *Reliability-Based Voting Schemes Using Modality-Independent Features in Multi-Classifier Biometric Authentication*, Proceedings of the 7^{th} International Workshop on Multiple Classifier Systems, Springer LNCS, vol. **4472**, pp. 377–386, 2007.

[16] TRAN D., *Estimation of Prior Probabilities in Speaker Recognition*, Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, pp. 141–144, 2004.