Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language 20 (2006) 192–209

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Using quality measures for multilevel speaker recognition

Daniel Garcia-Romero, Julian Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez *,
Javier Ortega-Garcia

ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior, Universidad Autonoma de Madrid,
Ctra. Colmenar km. 15 Campus de Cantoblanco, E-28049 Madrid, Spain

Received 2 November 2004; received in revised form 29 July 2005; accepted 31 August 2005
Available online 23 September 2005

Abstract

The use of quality information for multilevel speaker recognition systems is addressed in this contribution. From a definition of what constitutes a quality measure, two applications are proposed at different phases of the recognition process: scoring and multilevel fusion stages. The traditional likelihood scoring stage is further developed providing guidelines for the practical application of the proposed ideas. Conventional user-independent multilevel support vector machine (SVM) score fusion is also adapted for the inclusion of quality information in the fusion process. In particular, quality measures meeting three different goodness criteria: *SNR*, *F*0 deviations and the *ITU P*.563 objective speech quality assessment are used in the speaker recognition process. Experiments carried out in the Switchboard-I database assess the benefits of the proposed quality-guided recognition approach for both the score computation and score fusion stages.

1. Introduction

One of the key points addressed nowadays by automatic speaker recognition research is the exploitation of multilevel information in the speech signal (Reynolds et al., 2003; Campbell et al., 2003; Garcia-Romero et al., 2003). This idea is based on self-observation and experience, since listeners rely on several types or levels of information in the speech signal to recognize the speaker's identity. In the same way, it can be observed that humans are able to perform a number of sophisticated tasks, related to the quality of the information available and the sources of that information, when attempting to make a decision. For example, if a person is to make a decision about the identity of a speaker, based on a noisy and low fidelity speech recording, it is logical to think that the portions of the recording less corrupted by the noise should have a higher influence in the final decision. Furthermore, if the person has to make the decision based on the judgement of two experts, it is highly probable that the person will assign different credibilities to each expert depending on who they are or their previous experience.

[*] Corresponding author. Tel.: +34 914972269; fax: +34 914972235.
E-mail addresses: dgromero@atvs.diac.upm.es (D. Garcia-Romero), julian.fierrez@uam.es (J. Fierrez-Aguilar), joaquin.gonzalez@uam.es (J. Gonzalez-Rodriguez), javier.ortega@uam.es (J. Ortega-Garcia).

0885-2308/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2005.08.004

Based on these intuitive ideas and the functional structure of common speaker recognition systems, there are four potential stages in the recognition process in which the quality information may be incorporated (Garcia-Romero et al., 2004): feature extraction, model training, score computation and score fusion. Previous work in biometrics has shown promising results when incorporating quality measures into the recognition process (Bigun et al., 2003; Fierrez-Aguilar et al., 2005). In these studies, the quality information is incorporated at the score fusion level. The fusion function is adapted to rely more on the biometric traits that are less prone to error in noisy conditions. Other studies concerned with quality estimations in the field of speaker recognition include: (Koolwaaij et al., 2000) in which model quality assessment methods are studied to adapt the model training process and (Reynolds et al., 2003) in which quality-based feature selection is proposed to improve the performance of speaker recognition systems.

In addition to this previous work, new research efforts are also dedicated to the establishment of objective quality measures of biometric traits such as fingerprint (Simon-Zorita et al., 2003) and speech signals (Kim, 2004; ITU-T Recommendation P.563, 2004).

In this paper, we study the inclusion of quality information into speaker recognition systems by developing two applications at the scoring and multilevel fusion stages. On the one hand, the traditional likelihood scoring stage is extended to include the quality information in the score computation process. In particular, the score computation stage of a GMM-based speaker recognition system is adapted and tested. On the other hand, a conventional multilevel support vector machine (SVM) score fusion approach is adapted for the inclusion of quality information in the fusion process. The SVM quality-based score fusion paradigm described in Fierrez-Aguilar et al. (2005) is adapted to cope with the specificities of the multilevel speaker recognition process and integrated with the proposed automatic quality measures.

The remainder of this paper is structured as follows. In Section 2, we discuss the concept of quality measures and the inclusion of quality information into the speaker recognition process. In Section 3, we design three goodness criteria. In Section 4, we propose two novel applications for the quality measures at the score computation and fusion stages. In Section 5, we detail experimental results. Section 6 provides a summary of the main results and conclusions.

2. The concept of quality measure

The concept of quality may be defined[1] as the degree of goodness of an element given a certain criterion. This idea is quite similar to the underlying concept of a probability measure. Hence, to construct a mathematical model that quantifies the above notion, a quality measure function $Q^{\xi}(\cdot)$ may be formulated in probabilistic terms as follows:

$$Q^{\xi}(Y) = p(Y \text{ meets } \xi), \tag{1}$$

where ξ is a specific goodness criterion for the variable Y. As a result of this formulation, the quality measure function $Q^{\xi}(\cdot)$ assigns a number between 0 and 1 to every event, i.e., to every possible degree of goodness of Y given ξ. Hence, a reliable quality measure function should be able to quantify the quality of Y with a value of 1 when Y totally satisfies ξ and with a value of 0 when Y does not meet the established goodness criterion at all.

The crucial benefits brought into the recognition process by knowing the quality of the elements involved are significant, since this information allows the system to be dynamically adjusted. Examples include the importance given to certain portions of the incoming speech signal during the computation of its likelihood or even how the system relies on each of the scores produced by the different levels of information conveyed in the speech signal.

To some extent, there might be some confusion between the well-known concept of confidence measure, widely used in automatic speech recognition (ASR) (Siu et al., 1997), and the discussed idea of quality measure since both provide information that may be interpreted as how reliable a certain element involved in the recognition process is. It is important to notice that the essence of these two ideas is substantially different. The main purpose of a confidence measure is to indicate how correct is the estimated probability of a model matching some speech data (Williams, 1999), whereas the goal of a quality measure is to quantify how well a certain

[1] Cambridge Klett Dictionary.

goodness criterion is satisfied by an element of the system. Thus, the benefit of assigning a confidence estimate to a decoding is succinctly summarized by the phrase: "*knowing what you don't know*" (Williams, 1999), whereas the benefit of estimating the quality of an element of the recognition process is summarized by the phrase: "*knowing the quality of what you have*".

In order to incorporate this general concept of quality measure into the specific framework of speaker recognition systems, we can think of Y as any element of the system (e.g, speech signal, scores, models, thresholds, etc.) and ξ as any factor that affects the behavior of Y and hence the system performance (e.g, SNR, amount of data, course of time, etc.). For example, if we are working with data observed in noisy conditions, Y may be considered as the speech energy signal and ξ as a criterion based on SNR. Consequently, a quality measure may be stated as follows:

$$Q^{\xi=\mathrm{SNR}}(Y) = p(Y > \text{noise}). \tag{2}$$

If we consider the noise normally distributed with mean μ_t and variance σ_t, then the quality of the speech energy signal, $Y = \{y_t;\ t = 1, \dots ,T\}$, could be segmentally computed by means of the resulting expression (Renevey, 2000):

$$q_t^{\xi} = p(y_t > \text{noise}) = \int_{-\infty}^{y_t} \frac{1}{\sqrt{2\pi}|\sigma_t|} \exp\left(-\frac{(\theta - \mu_t)^2}{2\sigma_t^2}\right) \mathrm{d}\theta. \tag{3}$$

The resulting quality signal, $Q^{\xi} = \{q_t^{\xi};\ t = 1, \dots, T\}$, can be used by the speaker recognition system in several useful ways such as: eliminating the portions of the signal with low quality during the score computation or model training, incorporating the quality information in the score computation function, etc.

3. Goodness criteria

One of the key elements in obtaining a successful quality measure is the election of an adequate goodness criterion. Any factor affecting the behavior of an element of the speaker recognition system is susceptible to being used for the design of a goodness criterion.

It is useful to classify any goodness criterion based on its dependency or independency of the claimed identity. The reason for this is that identity-claim dependent goodness criteria need training information, related to the claimed identity, to be able to generate a quality signal. Moreover, this subset of criteria may have some speaker discriminative power since speaker information is used to train the criteria. However, on the other hand, identity-claim independent goodness criteria do not need any training information related to the claimed identity, and hence do not offer any discriminative power.

In the following, we are going to focus on three goodness criteria. The first one, $F0$ deviations from the mean, is identity-claim dependent, whereas the remaining two, SNR and ITU P.563, are identity-claim independent.

To perform the quality-based score computation (detailed in Section 4.1), it is necessary to obtain a quality signal at the frame level, whereas for the quality-based score fusion (described in Section 4.2) it is only necessary to have a quality value for the whole speech utterance. In the following section, we develop the proposed quality measures at the frame level, obtaining the overall quality of the speech utterances as the average of its corresponding quality signal.

3.1. *F0 deviations*

In order to design a goodness criterion, ξ_{F0} based on $F0$ deviations[2] from the mean, μ_{F0}, a model of the $F0$ distribution of the claimed identity is necessary. Due to the fact that the $F0$ distribution is Gaussian (Sonmez et al., 1997), the training speech of each user is used for the estimation of a user-dependent unimodal Gaussian model. For each test file, the quality value of each feature vector (belonging to a voiced region of the speech signal) is defined at discrete time instant t as

$$q_t^{\xi_{F0}} = p(|y_t^{F0} - \mu_{F0}| < |F0 - \mu_{F0}|), \tag{4}$$

[2] All $F0$ values are in a logarithmic scale.

where $F0 \sim N(\mu_{F0}, \sigma_{F0})$ is the pitch model of the claimed user, y_t^{F0} is the estimated pitch of the test segment at instant t and F_{F0} is the cumulative distribution function of $F0$.

For the unvoiced regions of the speech signal a fixed quality value, $q_{t_{\text{unv}}}$, is set a priori. In the following experiments, this value was heuristically set to 0.5.

3.2. SNR

During the design of a goodness criterion, ξ_{SNR}, based on SNR, the speech utterances were processed in segments of length between 5 and 20 s with a minimum voice activity ratio of 10%. These segments were obtained by dividing the speech utterance in the silence parts and merging or splitting them until they met the constraints mentioned above. The noise embedded in the speech signal was estimated during the silence parts of each speech segment k, being k the index of each speech segment. The noise was considered normally distributed with mean μ_k and variance σ_k. The quality of the speech energy signal, $Y = \{y_k;\ k = 1, \ldots, K\}$, was computed by means of the Eq. (3). The resulting quality signal, $Q^{\xi_{\text{SNR}}} = \{q_t^{\xi_{\text{SNR}}};\ t = 1, \ldots, T\}$, was obtained by assigning the same quality values to all the frames included in the speech segment being processed.

3.3. ITU P.563 objective speech quality assessment

Due to the fact that new research efforts are dedicated to the establishment of objective quality measures of speech signals, it is interesting to assess their performance for the specific purpose of aiding speaker recognition systems. For this reason, we considered the ITU-T P.563 recommendation (ITU-T Recommendation P.563, 2004), which describes an objective single-ended method for predicting the subjective quality of telephonic speech signals. The P.563 approach is the first recommended method for single-ended non-intrusive measurement applications that takes into account the full range of distortions occurring in switched telephone networks and that is able to predict the speech quality on a perception based scale MOS-LQO (ITU-T Recommendation P.800.1, 2003). Since the output of the P.563 system is a value in the range [1,5] estimating the perceived quality of the speech utterance, it is very reasonable to consider this system as an "electronic ear" that quantifies the quality of the speech signal regarding factors such as echo, noise, channel errors, etc.

In order to use the information provided by this approach as a goodness criterion $\xi_{P.563}$, we made the output of the P.563 system to be compliant with our quality measure definition, by linearly mapping it into a [0,1] range. Following the P.563 recommendation, the speech utterances were processed in segments of length between 5 and 20 s with a minimum voice activity ratio of 10%. These segments were obtained by dividing the speech utterance between the silence parts and merging or splitting them until they were compliant with the constraints mentioned above. By doing this, we obtained a temporary quality signal $Q_{\text{temp}}^{\xi_{P.563}} = \{q_k^{\xi_{P.563}};\ k = 1, \ldots, K\}$, where k was the segment index. The final quality signal, $Q^{\xi_{P.563}} = \{q_t^{\xi_{P.563}};\ t = 1, \ldots, T\}$, was obtained by assigning the same quality values to all the frames included in the speech segment being processed. Fig. 1 shows an example of a quality measure, based on P.563, in which it is easy to notice how all the frames in the same speech segments were assigned the same quality value.

4. Application of quality measures

There are four potential stages for the inclusion of the quality information in the recognition process (Garcia-Romero et al., 2004): feature extraction, model training, score computation and score fusion. In the present work, we focus on the score computation and score fusion stages. See Fig. 2 for the general system model.

4.1. Quality-based score computation

The state of the art in speaker recognition systems has been widely dominated during the past decade by the UBM-MAP adapted GMM approach working at the short-time spectral level (Reynolds et al., 2000). Recently, new approaches based on support vector machines (SVM) (Campbell, 2002) are achieving similar performance, working at the spectral level, and also providing complementary information useful for the fusion of both approaches (Campbell et al., 2004). Furthermore, higher levels of information conveyed in the speech

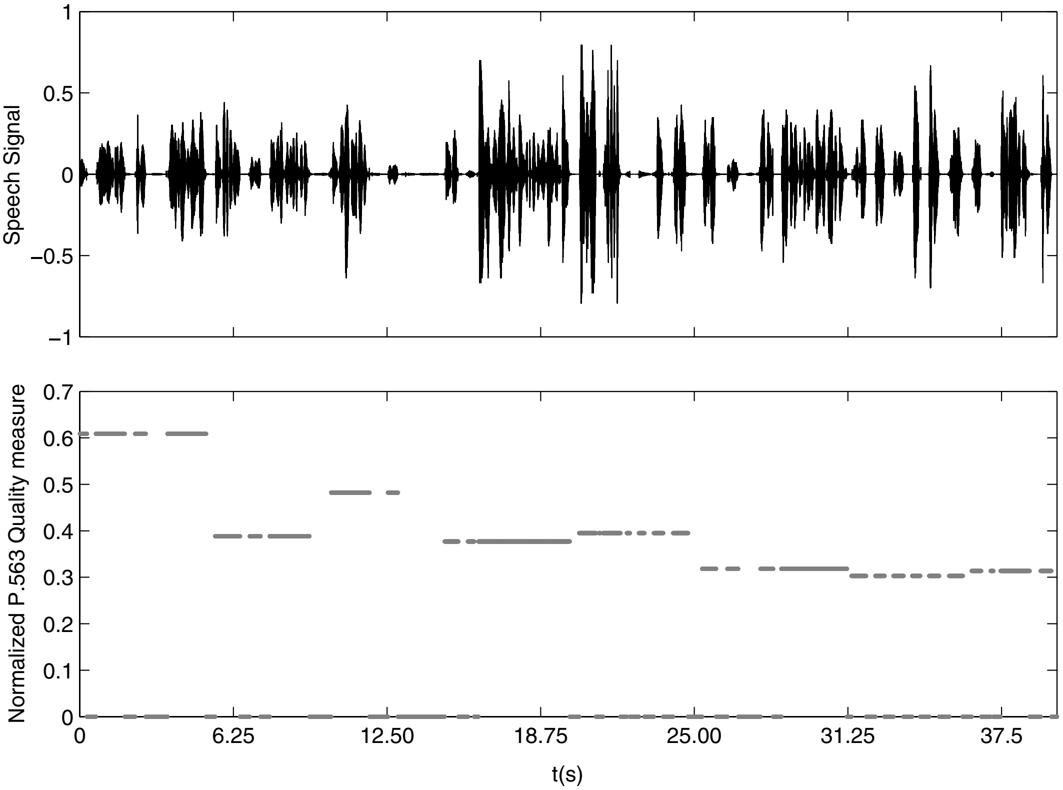

Fig. 1. Example of a normalized *P*.563 quality signal (bottom) for a SWB-I speech segment (top).

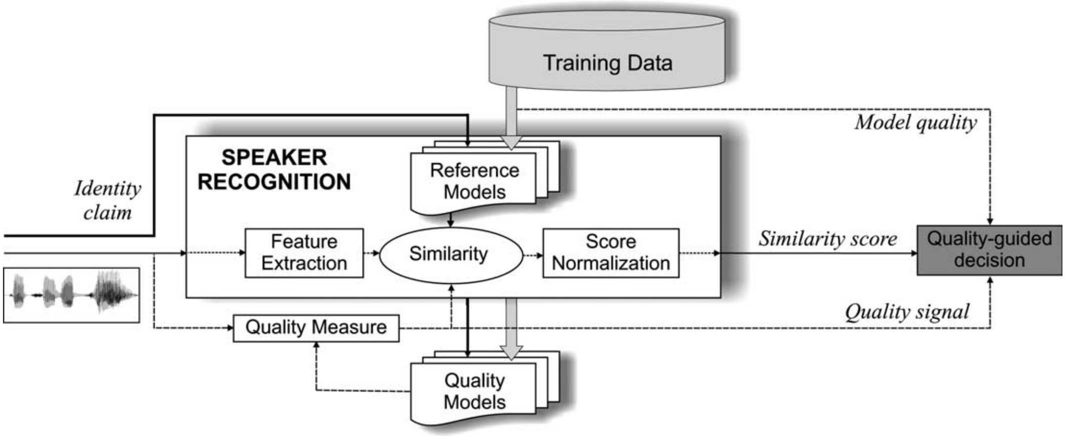

Fig. 2. General system model for speaker recognition using quality measures.

signal have shown promising discriminative capabilities among speakers and are a major goal of present Speaker Recognition research efforts (Reynolds et al., 2003).

A common practice shared among all the above-mentioned Speaker Recognition techniques is the use of a pre-processing stage in which two major tasks are accomplished: (i) the signal is enhanced according to certain criteria (e.g., channel effects removal, noise reduction, etc.); (ii) hard decisions about the correctness of the basic constituting elements of the data are made (e.g., silence removal, non-speech sound rejection, etc.), preserving those pieces of information that satisfy certain criteria and dismissing the remaining ones.

This pre-processing approach, combined with a conventional scoring mechanism, has the drawback of regarding all the preserved information as equal in terms of importance once the signal has been pre-processed. Therefore it omits, during the score computation process, the fact that both the information concerning speaker identity and the perturbing artifacts are not distributed uniformly along the pre-processed signal (Malayath et al., 2000).

Previous work in speech recognition has shown that some speech features are more important than others depending on the phonetic context (Rogina and Waibel, 1994). To take this fact into account, individual stream weights were learned for each HMM state and included in the class-dependent probability estimation process. In a similar way, the underlying idea in the quality-based score computation (QBSC) approach suggests the incorporation of estimated quality measures (carried out during pre-processing) as weighting factors in the score computation process.

The QBSC concept is applicable to any of the aforementioned techniques used in Speaker Recognition systems. In the following, we are going to particularize for the case of GMM's working at the short-term spectral level, since it is the most widely used paradigm for speaker recognition (Reynolds, 2002).

4.1.1. Quality-based GMM score computation

For a D-dimensional feature vector, $\mathbf{o}$, and a weighted linear combination of M unimodal Gaussian densities, $p(\mathbf{o}|\lambda)$, with the parameters of the density model denoted as

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, \ldots, M \tag{5}$$

the likelihood function is defined as

$$p(\mathbf{o}|\lambda) = \sum_{i=1}^{M} w_i p(\mathbf{o}|\mu_i, \Sigma_i). \tag{6}$$

Given a sequence of feature vectors, $O = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$, usually assumed independent, and a quality signal

$$Q^{\xi} = \{q_1^{\xi}, q_2^{\xi}, \ldots, q_T^{\xi}\} \tag{7}$$

computed through the speech signal Y with a specific goodness criterion ξ, the likelihood of the model λ incorporating the quality measure as a weighting factor is denoted

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

The log-likelihood is computed as

$$\log p(O|Q, \lambda) = \sum_{t=1}^{T} q_t^{\xi} \log p(\mathbf{o}_t|\lambda) \tag{9}$$

Often, the average log-likelihood is used to normalize out duration effects from the likelihood value. This can be accomplished by dividing Eq. (9) by $\sum_{t=1}^{T} q_t^{\xi}$. Since the assumption of independence between the feature vectors is not precise, this scaling factor can be consider as a rough duration compensation (Reynolds et al., 2000).

If a quality measure that works in spectro-temporal regions (i.e., assigns quality values to each feature vector coefficient) is used instead of one that works in temporal regions (same quality assigned to the entire feature vector), conventional missing data approaches, such as bounded marginalization (BMG) or bounded data imputation (BDI), can be used for the likelihood computation (Barker et al., 2000).

4.2. Quality-based score fusion

In order to exploit the different levels of information conveyed in the speech signal (e.g., lexical, phonetic, spectral, etc.) (Campbell et al., 2003; Garcia-Romero et al. , 2003) efficient score combination methodologies are necessary (Reynolds et al., 2003). This problem can be formulated as the fusion of different machine experts.

Two theoretical frameworks for combining classifier outputs with application to biometric authentication are described in Bigun et al. (1997) and Kittler et al. (1998). The former is derived from a risk analysis perspective (Bigun, 1995) and the later is based on statistical pattern recognition theory (Duda et al., 2001). Both of them concluded (under some mild conditions which deserve further attention (Kittler and Alkoot, 2003) that the weighted average is a good way of conciliating the confidences (similarity scores) provided by the different recognition systems involved.

Interestingly enough, the approach in Bigun et al. (1997) was further developed in Bigun et al. (2003) providing guidelines for the use of quality measures in combining classifiers. In particular, a quality-based score fusion scheme was derived in which the output of the classifiers was adapted based on the estimated quality of the input traits. The fusion function was adapted to rely more on the traits that were less prone to error in noisy conditions. This basic idea has also been recently exploited using discriminative learning approaches (Toh et al., 2004; Fierrez-Aguilar et al., 2005). In Toh et al. (2004), polynomial decision functions were used for combining classifiers and some quality measures were included as regularization terms in the discriminative training process. In Fierrez-Aguilar et al. (2005), SVM decision functions were used for combining classifiers and quality measures were used as trade-off coefficients between different decision functions.

In the following, we address the specificities of the quality-based score fusion (QBSF) applied to multilevel speaker recognition. We also propose an operational QBSF scheme using SVMs (Fierrez-Aguilar et al., 2005) that is adapted for multilevel speaker recognition systems.

4.2.1. Quality-based score fusion for multilevel speaker recognition

In order to present a clear study of the inclusion of quality information in the score fusion stage, we are going to focus on the combination of low-level speaker information (i.e., spectral information) with two high-level sources of speaker information (i.e., phonetic and lexical). Furthermore, we are only going to study the particular case of combining spectral information with each of the mentioned high-level information sources, hence yielding a two-level combination.

Fig. 3 shows the proposed QBSF model for this particular case of study. The design of this model relies on the following premises: (i) up to date, Speaker Recognition systems based on low-level information (i.e., spectral information) achieve better performance than individual high-level information systems (Reynolds et al., 2003; Campbell, 2003), and (ii) artifacts degrading the performance of low-level recognition systems are better identified and studied than those affecting high-level recognition systems, hence making the design of quality measures for low-level systems easier than those for the high-level ones.

Based on the scenario described above, we propose a QBSF approach in which the quality information is incorporated as a trade-off between (a) only using the recognition system with the best performance (i.e., low-level system) and (b) the combination of both systems (i.e., low-level and high-level).

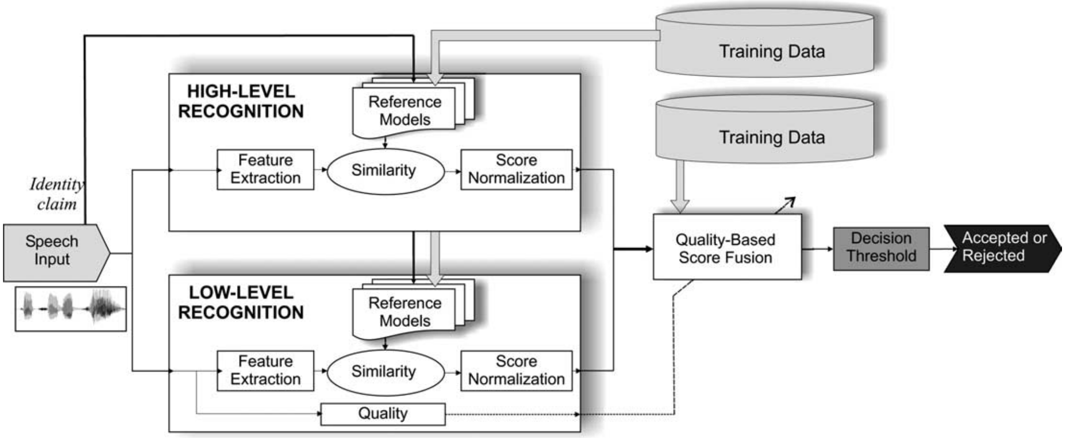

Fig. 3. General system model for multilevel speaker recognition using quality measures.

The proposed multilevel speaker recognition QBSF approach has been adapted from the previously developed multimodal QBSF scheme described in Fierrez-Aguilar et al. (2005). The difference lies in the terms involved in the quality-based trade-off. Whereas in the previous multimodal case, any trait with low quality values could be completely discarded; in the multilevel speaker recognition case the low-level information system is never completely discarded. This is due to the fact that low-level speaker information yields much better results than any current high-level information system. In this way, the quality value of the speech segment determines if the final score is computed based only on the low-level information or on both low- and high-level information. This change in the system model ensures that the fused score is at least as good as the score of the best performing system (i.e., low-level system) if not better. If in the future, the systems using high-level information achieve similar performances to those using low-level information, the previous multimodal QBSF scheme (Fierrez-Aguilar et al., 2005) will also be applicable to the speaker recognition case.

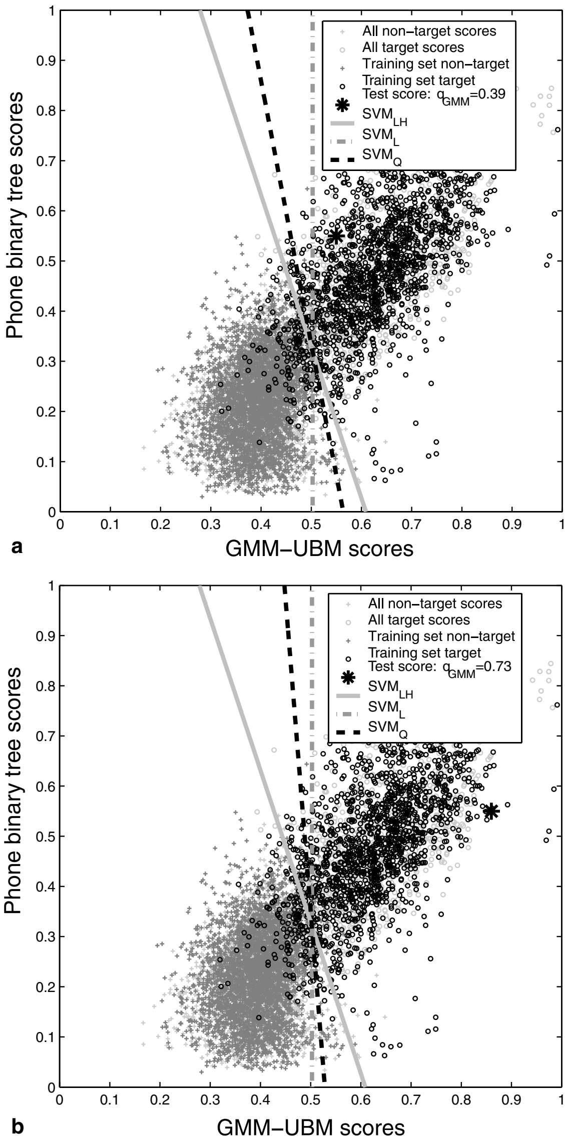

Fig. 4. SVM_Q fusion hyperplane as a trade-off between SVM_L and SVM_{LH} for two multilevel scores with different quality values: (a) $q_{GMM} = 0.39$ and (b) $q_{GMM} = 0.73$.

4.2.2. Quality-based SVM score fusion

Given R speaker recognition systems working with different information levels, each one computes a similarity score x_r, $r = 1, \dots, R$, between the test speech signal S and the speaker model. Let the individual similarity scores be combined into a multilevel score $\mathbf{x} = [x_1, \dots, x_R]^{\mathrm{T}}$. In our particular case of study, the multilevel score is the combination of two individual scores $\mathbf{x} = [x_{\mathrm{L}}, x_{\mathrm{H}}]^{\mathrm{T}}$, where x_{L} and x_{H} are the scores of the low- and high-level systems, respectively. Let q_{L} denote the quality signal obtained from a quality measure (i.e., SNR) of the test speech signal S affecting the low-level recognition system. The proposed operational QBSF scheme (from now on also referred to as $\mathrm{SVM_Q}$) is as follows:

(1) *SVM_Q training*: Given a labeled training set of N_{TR} multilevel scores, $(\mathbf{x}_i, y_i)$, with $y_i \in \{-1, 1\} = \{\text{Impostor, Client}\}$, and $i = 1, \dots, N_{\mathrm{TR}}$ indexing the training sample. A linear SVM-based fusion scheme of the low- and high-level systems ($\mathrm{SVM_{LH}}$) is trained using standard procedures (Fierrez-Aguilar et al., 2004), but computing the costs coefficients, C_i, of each training sample of the SVM regularized cost function as follows:

$$C_i = C \cdot q_{i,\mathrm{L}}. \tag{10}$$

In this way, C_i is the product between the quality information, $q_{i,\mathrm{L}}$, associated with the training score $\mathbf{x}_i$ and a positive constant C. As a result, the higher the quality of the speech used for generating the training score, the higher its contribution in training the fusion function. Additionally, another SVM of dimension one ($\mathrm{SVM_L}$) is trained by using the training data from the low-level (spectral) system and the coefficients in Eq. (10).

(2) *SVM_Q authentication phase*: At this step, the two classifiers mentioned above, $\mathrm{SVM_{LH}}$ and $\mathrm{SVM_L}$, are already trained. When an input speech segment S, with its quality measure q_{L} is available, along with a claimed identity, the system generates a multilevel similarity score $\mathbf{x} = [x_{\mathrm{L}}, x_{\mathrm{H}}]^{\mathrm{T}}$. Finally, the combined quality-based similarity score is computed as follows:

$$f_{\mathrm{SVM_Q}}(\mathbf{x}) = q_{\mathrm{L}} f_{\mathrm{SVM_L}}(x_{\mathrm{L}}) + (1 - q_{\mathrm{L}}) f_{\mathrm{SVM_{LH}}}(\mathbf{x}), \tag{11}$$

where $f_{\mathrm{SVM_L}}(\cdot)$ and $f_{\mathrm{SVM_{LH}}}(\cdot)$ are signed distances to the linear decision hyperplanes provided by $\mathrm{SVM_L}$ and $\mathrm{SVM_{LH}}$, respectively (Fierrez-Aguilar et al., 2004).

Fig. 4 shows an example of a $\mathrm{SVM_Q}$ linear decision hyperplane computed as a trade-off between $\mathrm{SVM_L}$ and $\mathrm{SVM_{LH}}$ for two multilevel scores with different quality values. As indicated in Eq. (11), the higher the quality value, the higher the contribution of the low-level system to the final fusion function.

5. Experiments

5.1. Switchboard I database

Partitions 1, 2 and 3 of the Switchboard I database (SWB-I), as defined in Reynolds et al. (2003), have been used for the performance assessment of the proposed quality-based approaches on landline telephone data. The number of speaker models involved is 486 (260 male + 226 female). Each target model has been trained with a speech segment of approximately 2.5 min comprising one side of a 5-min telephonic conversation. Two different test sets have been used for the system assessment: (i) one side of the conversation test segments (approx. 2.5 min. of speech); (ii) two sides of the conversation test segments (approx. 5 min. of speech). The total number of trials obtained with each test set is 8248 (2416 target, 5832 non-target). For the QBSC experiments, both test sets were used, whereas for the QBSF experiments only the one side test segment subset was used.

5.2. Baseline systems description

Three different speaker information levels have been selected for the experiments in this paper: spectral, phonetic and lexical. The spectral level is used in all the experiments, whereas the phonetic and lexical levels are only used for the QBSF experiments.

The spectral level has been selected since it is the speaker information level that, up to date, has the best performance for speaker recognition (Campbell, 2003). The phonetic level was selected because it is the high-level speaker information source in which the closest performance to the spectral level has been reported for the Switchboard I database (Reynolds et al., 2003). Finally, the lexical level was selected to assess the performance of the proposed QBSF between two systems, spectral and lexical, with very different performances.

Now we are going to give a brief description of the three baseline systems and provide references for further details.

5.2.1. Spectral system

A UBM-MAP adapted GMM system (Garcia-Romero et al., 2004) with 256 mixtures and diagonal covariance matrix was used to model the feature vectors (19 MFFC + 19 ΔMFCC) obtained every 10 ms with a 20-ms Hamming window. The score computation was performed as a likelihood ratio (LR) between the target model and the UBM likelihoods.

The resulting scores provide a baseline result for comparison with the proposed QBSC-GMM system. Fig. 5 shows the performance of the spectral system for the Switchboard I database, obtaining a 6.13% of Equal Error Rate (EER).

5.2.2. Phonetic system

Capturing speaker-dependent pronunciation by means of modelling phone sequences has shown to be a viable and effective approach to speaker recognition (Andrews et al., 2002). The phonetic system providing the scores for the present work was developed at the SuperSID project (Reynolds et al., 2003) and uses a binary-tree-structured statistical model for extending the phonetic context beyond of standard n-grams without exponentially increasing the model complexity (Klusacek et al., 2003). Fig. 5 shows the performance of the phonetic system for the Switchboard I database, obtaining a 11.60% EER.

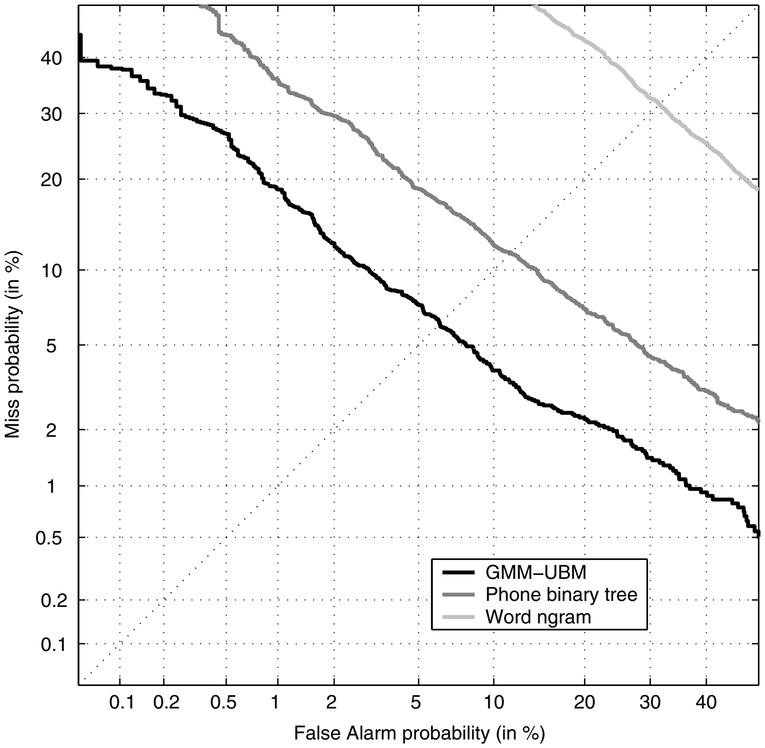

Fig. 5. DET curves of the baseline systems for the Switchboard I database.

5.2.3. Lexical system

The *n*-gram (bigram) idiolect system providing the scores for the present work was developed in the SuperSID project (Reynolds et al., 2003) and is based on an LR computation between the target and background model likelihoods. The LR computation was performed following the procedures developed in (Doddington, 2001). The word transcripts used for the score computation were provided by the ASR Dragon system with an 11% of Word Error Rate (WER). Fig. 5 shows the performance of the lexical system for the Switchboard I database, obtaining a 31.60% EER.

5.3. Quality-based score computation experiments

In Fig. 6, the performance assessment of both the baseline GMM Speaker Verification system and the QBSC adaptation, with three different quality measures, are depicted in the form of DET plots for the above described corpora.

In relation to the one-side test set of the SWB-I database, see Fig. 6, a slight improvement is obtained by means of using the QBSC adaptation of the baseline system for all the quality-measures. This result is more noticeable in the low false alarm and low miss probability regions of the DET curve. Table 1 shows the baseline and QBSC adaptation performance for the EER operational point with the three quality measures. The analysis of the EER values for the male and female partitions show that the three quality measures have a similar behavior across genders.

A bigger improvement is obtained for the test set comprising both sides of the telephonic conversation, see Fig. 6 and Table 2. The fact that 2 speakers are involved in the test segments makes this set more suitable for the achievement of better results since a larger portion of the speech signal is considered corrupted. In the special case of a quality measure capable of quantifying the speech segments not belonging to the target speaker with a low quality value, the QBSC adapted system may perform some kind of "speaker spotting". This may be the case for the selected *F*0-based quality criterion since it is possible to discriminate among speakers based on *F*0 information (Sonmez et al., 1997). Therefore, the "speaker spotting" effect of the selected *F*0 quality measure provides a justification for the better performance on the test set comprising both sides of the telephonic conversation. The performance improvement for the *SNR* and *P*.563 goodness criteria is lower than the obtained by *F*0-based quality criterion since these criteria are not able to perform the "speaker spotting" effect mentioned above. It should be clear that any speaker discriminant quality measures (not just *F*0) will also lead to a performance improvement for the specific case of test segments with more than one speaker.

5.4. Quality-based score fusion experiments

5.4.1. Experimental protocol

In order to perform a fair assessment of the fusion approaches, it is necessary to split the available scores into training and testing sets. The reason for that is to avoid assessing the fusion system with scores used during the training phase of the fusion function. A threefold cross-validation approach, using partitions 1, 2 and 3 of the SWB-I database, was designed for that purpose. All the scores within a partition were obtained using speaker models and test segments from within that partition only. In this way, when the scores of one partition are used for testing the fusion rule, the scores of the remaining two partitions are used for training the fusion function. As mentioned in Section 4.2.1, we are going to focus on two-level system fusion: spectral-phonetic and spectral-lexical.

5.4.2. Results

To obtain a better understanding of the fundamental principles supporting the use of quality measures in the fusion process, Fig. 7 shows a scatter plot of the scores of the baseline systems and their corresponding *F*0 and *P*.563 quality values. The *SNR* quality measure is not depicted since its behavior is very similar to the *P*.563 scatters. Linear regression fits were computed separately for target and non-target scores in each of the scatter plots.

It is desirable that a quality measure meets the following properties: (i) non-target scores and their corresponding quality values should have a negative correlation so the bigger the quality value the smaller the

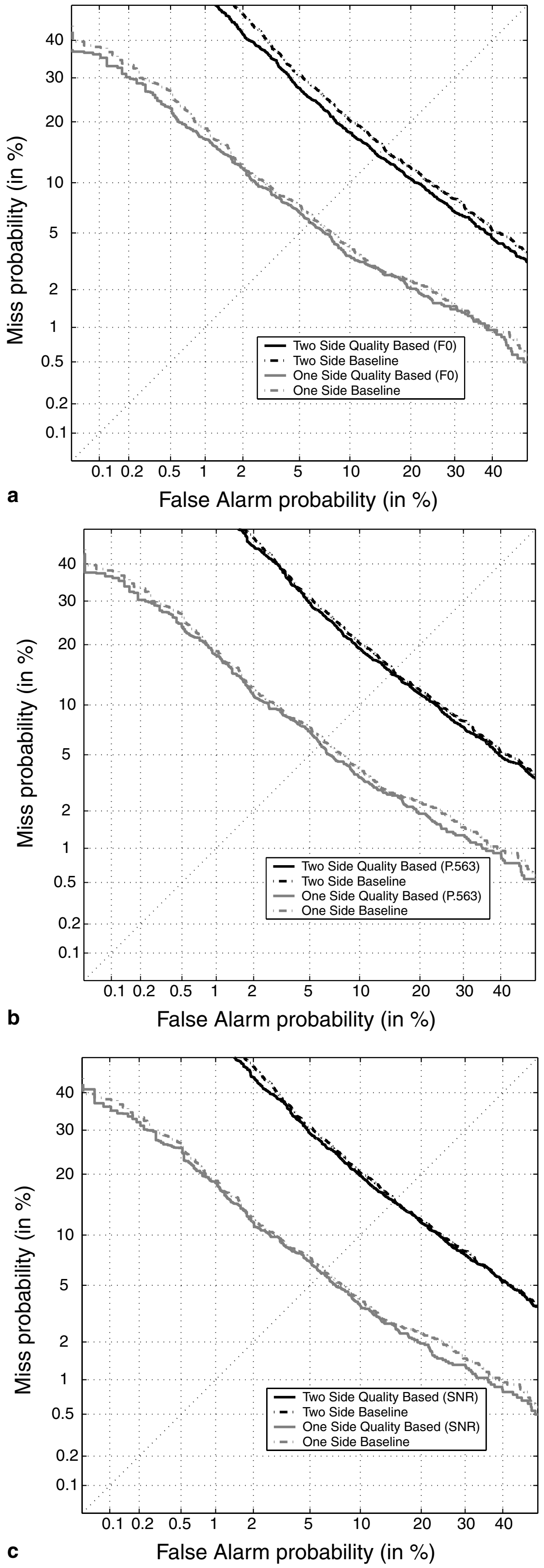

Fig. 6. System performance on Switchboard I database with: (a) *F*0, (b) *P*.563 and (c) *SNR* quality measures.

Table 1
Results on the Switchboard I one-side test set

Partition	EER (%)			
	Baseline	Q-based(F0)	Q-based(P.563)	Q-based(SNR)
Female	5.64	5.31	5.53	5.38
Male	6.31	6.09	6.03	6.03
Pooled	6.13	5.88	6.05	5.88

Table 2
Results on the Switchboard I test set comprising both sides of the conversation

Partition	EER (%)			
	Baseline	Q-based(F0)	Q-based(P.563)	Q-based(SNR)
Female	15.52	14.52	15.56	15.41
Male	14.72	13.34	14.49	14.36
Pooled	15.09	14.01	14.95	14.73

score; (ii) target scores and their corresponding quality values should have a positive correlation so the bigger the quality value the bigger the score.

Keeping these properties in mind and observing the slopes of the linear regressions, two major ideas are worth noting: (i) the behavior of the three quality measures tend to be better for the low-level (spectral) system; (ii) the higher the information level (lexical > phonetic > spectral) the less affected it is by the disturbing artifacts considered by the three quality measures. This asseveration is in accordance with previous research results (Campbell, 2003), and it is one of the main motivations for using high-level speaker information systems. Finally, recalling the classification of the goodness criteria established in Section 3, the first row in Fig. 7 shows how identity-claim dependent goodness criteria (such as ξ_{F0}) may have some speaker discriminative power in themselves.

In order to compare the SVM-QBSF approach with the standard SVM fusion approach (Fierrez-Aguilar et al., 2004), we carried out a series of experiments comparing the system performances, in terms of EER, as a function of the number of models used for training. Each experiment was conducted following the procedures described in the experimental protocol and sampling the corresponding training sets using 4 bootstrap iterations in which the scores of M models, with $M \in \{10, 20, 40, 80, 120, 162\}$, were randomly chosen without replacement. Fig. 8 shows the results for the spectral-phonetic and the spectral-lexical fusion systems. It is worth pointing out that the SVM-QBSF approach is less sensitive to the number of models (amount of data) in the training set than the standard SVM approach. In this sense, the quality information may be helping the quality-based fusion system to generalize better than the standard SVM approach. Moreover, the performance of the SVM-QBSF approach is better than the performance of the standard SVM approach for a number of models, $M < 120$, in both the spectral-phonetic and the spectral-lexical fusion systems.

Fig. 9 shows the performance of spectral-phonetic and the spectral-lexical fusion systems with $F0$ quality measures for the particular case of $M = 20$. The performance for the SNR and $P.563$ quality measures are very similar to the $F0$ results. For both systems, the SVM-QBSF approach outperforms the standard SVM approach in all the operating points for each of the quality measures. It is interesting to realize that for the spectral-lexical fusion systems the standard SVM approach obtains a performance worse than the individual spectral system. This result may be caused by a poor generalization of the fusion approach based on a small training set. Hence, the SVM-QBSF approach may be a good alternative for applications in which large training data sets are not available or there is a severe mismatch between development and testing data.

In general, the three quality measures reveal similar trends in terms of performance improvement of the fusion system. Table 3 shows that the correlation coefficients between each pair of quality measure are considerably small. The correlation coefficient between the SNR and the $P.563$ criteria is the highest. This may be due to the fact that both quality measures are computed following the same segmentation strategy

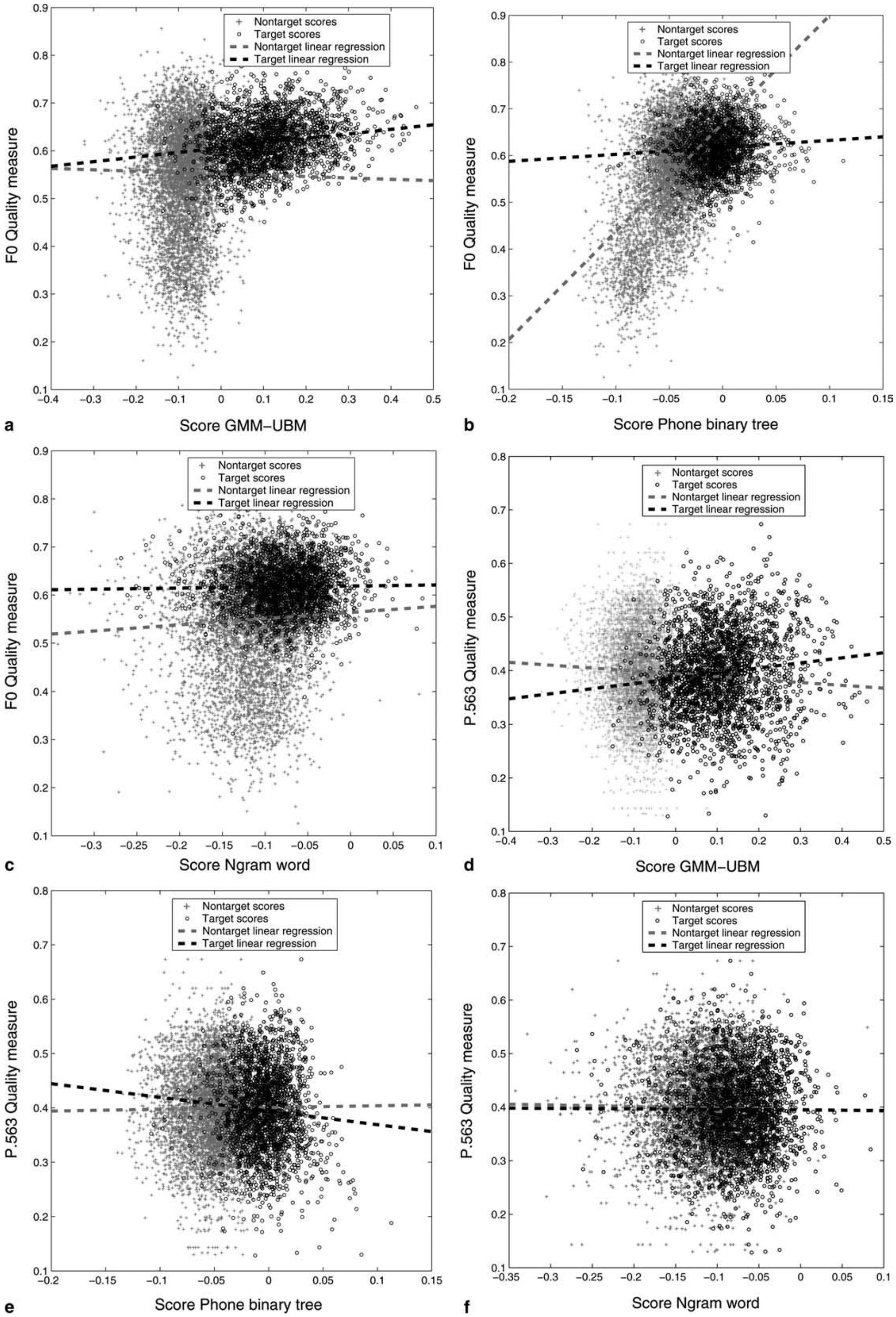

Fig. 7. Scatter plot of the *F*0 (a–c) and *P*.563 (d–f) quality measure values vs. the GMM-UBM, Phone-Binary Tree and *n*-gram word scores.

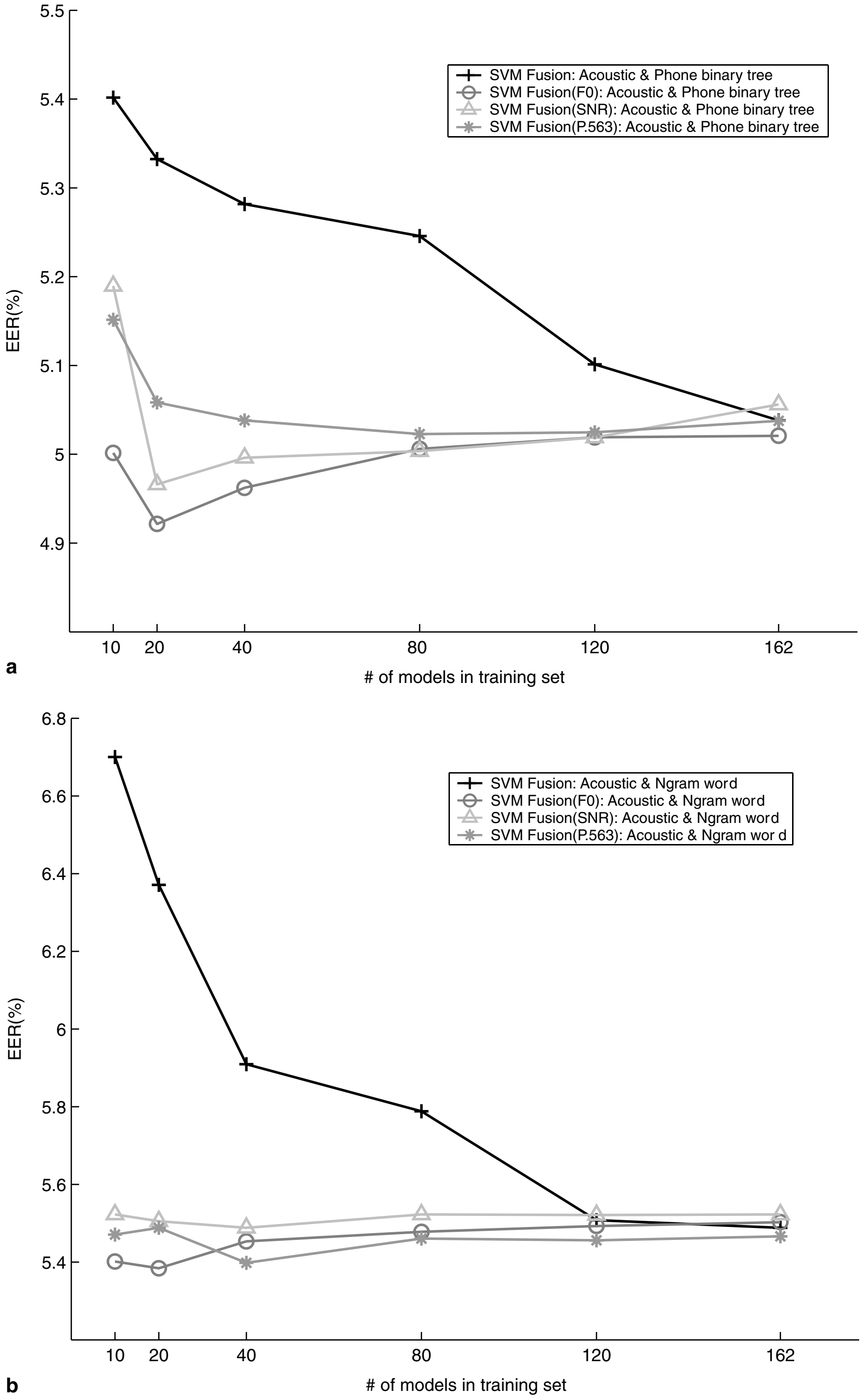

Fig. 8. Effect of the number of training models in the SVM fusion: (a) spectral-phonetic information and (b) spectral-lexical information.

(see Section 3 for more details). The fact that the correlation coefficients are small and the improvement yielded by each quality measure is similar, suggest that the combination of the three quality measures may be a good practice.

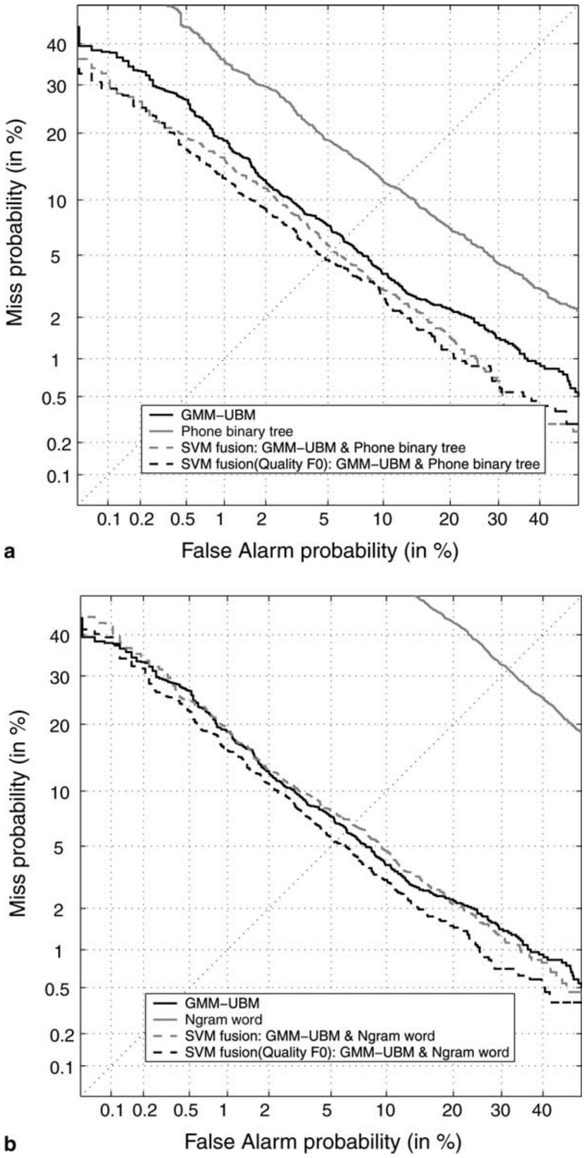

Fig. 9. SVM fusion results of: (a) spectral-phonetic and (b) spectral-lexical systems for the *F*0 quality measure.

Table 3
Correlation coefficients for each pair of quality measures

Quality measures	*F*0–*SNR*	*F*0–*P*.563	*SNR*–*P*.563
Correlation coefficient	0.18	0.35	0.46

6. Conclusions

An overview of the use of quality information for automatic speaker recognition systems has been reported. Two quality-based applications, at different phases of the recognition process, have also been proposed: Quality-based Score Computation and Quality-based Score Fusion. In the former, traditional likelihood scoring of a GMM has been further developed providing guidelines for the practical application of the proposed ideas. In the latter, standard SVM fusion approach has been adapted to take into account the quality information of the input speech. Experiments carried out on QBSC corroborate the benefits of the proposed quality-guided recognition approach on landline data for different quality measures. In particular, three frame-level quality measures meeting goodness criteria based on: *F*0 deviations, *SNR* and *ITU P*.563 recommendation have been used. Up to 7.15% of relative improvement at the EER operational point has been obtained on the Switchboard-I database. Experiments performed on SVM-QBSF have proved this approach to be less sensitive to the amount of training data than the standard SVM approach, hence demonstrating SVM-QBSF to be a robust fusion scheme for applications in which large data sets are not available for training or there is a severe mismatch between development and testing data.

Acknowledgements

This work has been supported by the Spanish Ministry for Science and Technology under projects TIC2003-09068-C02-01 and TIC2003-08382-C05-01. J. F.-A. thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research. The authors thank Douglas Reynolds for providing the SuperSID scores used in the QBSF experiments.

References

Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., Hernandez-Cordero, J., 2002. Gender-dependent phonetic refraction for speaker recognition. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 149–152.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: International Conference on Speech and Language Processing, ICSLP, pp. 538–561.

Bigun, E.S., Bigun, J., Duc, B., Fischer, S., 1997. Expert conciliation for multi modal person authentication systems by Bayesian statistics. In: Bigun, J., Chollet, G., Borgefors, G. (Eds.), Audio and Video based Person Authentication, AVBPA97 . Springer, pp. 291–300.

Bigun, J., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., 2003. Multimodal biometric authentication using quality signals in mobile communications. In: Proceedings of IAPR International Conference on Image Analysis and Processing, ICIAP. IEEE Computer Society Press, pp. 2–13.

Bigun, E.S., 1995. Risk analysis of catastrophes using experts' judgments: An empirical study on risk analysis of major civil aircraft accidents in Europe. European Journal of Operational Research, 599–612.

Campbell, J.P., Reynolds, D.A., Dunn, R.B., 2003. Fusing high- and low-level features for speaker recognition. In: Proceedings of the Eurospeech, pp. 2665–2668.

Campbell, W.M., Reynolds, D.A., Campbell, J., 2004. Fusing discriminative and generative methods for speaker recognition: Experiments on Switchboard and NFI/TNO field data. In: The Speaker and Language Recognition Workshop, Odyssey, pp. 41–44.

Campbell, W.M., 2002. Generalized linear discriminant sequence kernels for speaker recognition. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 161–164.

Campbell, J.P., 2003. Advances in speaker recognition: Getting to know you, Biometric Consortium Conference, MIT Presentation.

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. In: Proceedings of the Eurospeech, pp. 2521–2524.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley.

Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J., 2004. Exploiting general knowledge in user-dependent fusion strategies for multimodal biometric verification. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 617–620.

Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J., 2005. Discriminative multimodal biometric authentication based on quality measures. Pattern Recognition, 777–779.

Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2003. Support vector machine fusion of idiolectal and acoustic speaker information in spanish conversational speech. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 229–232.

Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2004. On the use of quality measures for text-independent speaker recognition. In: The Speaker and Language Recognition Workshop, Odyssey, pp. 105–110.

Garcia-Romero, D., Ramos-Castro, D., Gonzalez-Rodriguez, J., 2004. ATVS-UPM results and presentation at NIST'2004 speaker recognition evaluation (July 2004).

ITU-T Recommendation P.563. 2004. Single ended method for objective speech quality assessment in narrow-band telephony applications.

ITU-T Recommendation P.800.1, 2003. Mean opinion score (MOS) terminology.

Kim, D.-S., 2004. Special session: objective quality assessment of speech. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP.

Kittler, J., Alkoot, F.M., 2003. Sum versus vote fusion in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 110–115.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 226–239.

Klusacek, D., Navratil, J., Reynold, D.A., Campbell, J.P., 2003. Conditional pronunciation modeling in speaker detection. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 804–807.

Koolwaaij, J., Boves, L., Jongebloed, H., den Os, E., 2000. On model quality and evaluation in speaker verification. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 3759–3762.

Malayath, N., Hermansky, H., Kajarekar, S., Yegnanarayana, B., 2000. Data-driven temporal filters and alternatives to GMM in speaker verification. Digital Signal Processing, 55–74.

Renevey, P., 2000. Speech recognition in noisy conditions using missing feature approach. Ph.D. thesis, Ecole Polytechnique Federale de Lausanne.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing, 19–41.

Reynolds, D.A. et al. 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 784–787.

Reynolds, D.A., 2002. An overview of automatic speaker recognition technology. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 4072–4075.

Rogina, I., Waibel, A., 1994. Learning state-dependent stream weights for multi-codebook HMM speech recognition systems, International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 217–220.

Simon-Zorita, D., Ortega-Garcia, J., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., 2003. Image quality and position variability assessment in minute-based fingerprint verification, IEE Proceedings Vision, Image and Signal Processing, 150 (6), 402–408 (special issue on Biometrics on the Internet).

Siu, M., Gish, H., Richardson, F., 1997. Improved estimation, evaluation and applications of confidence measures for speech recognition. In: Fifth European Conference on Speech Communication and Technology, EuroSpeech, pp. 831–834.

Sonmez, M.K., Heck, L., Weintraub, M., Shriberg, E., 1997. A lognormal tied mixture model of pitch for prosody-based speaker recognition. In: Fifth European Conference on Speech Communication and Technology, EuroSpeech.

Toh, K.-A., Yau, W.-Y., Lim, E., Chen, L., Ng, C.-H., 2004. Fusion of auxiliary information for multi-modal biometrics authentication. In: Proceedings of the International Conference on Biometric Authentication, ICBA, Lecture Notes in Computer Science, vol. 3072, pp. 678–685.

Williams, D., 1999. Knowing what you don't know: roles for confidence measures in automatic speech recognition. Ph.D. thesis, University of Sheffield, England.