Available online at www.sciencedirect.com
ScienceDirect
Journal of the Franklin Institute 343 (2006) 574–595
Journal of The Franklin Institute
www.elsevier.com/locate/jfranklin

Confidence and reliability measures in speaker verification

Jonas Richiardi*, Andrzej Drygajlo, Plamen Prodanov

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne, EPFL-STI-ITS-LIDIAP, ELD 243, Station 11, 1015 Lausanne, Switzerland

Received 29 December 2005; received in revised form 7 July 2006; accepted 13 July 2006

Abstract

Speaker verification is a biometric identity verification technique whose performance can be severely degraded by the presence of noise. Using a coherent notation, we reformulate and review several methods which have been proposed to quantify the uncertainty in verification results, some with a view to coping with the effects of mismatched training-testing environments. We also include a recently proposed method, which is firmly rooted in a probabilistic approach and interpretation, and explicitly measures signal quality before assigning a reliability value to the speaker verification classifier's decision. We evaluate the performance of the confidence and reliability measures over a noisy 251-users database, showing that taking into account signal-domain quality can lead to better accuracy in prediction of classifier errors. We discuss possible strategies for using the measures in a speaker verification system, balancing acquisition duration and verification error rate.

Keywords: Confidence estimation; Reliability estimation; Error modelling; Speaker verification; Bayesian networks

1. Introduction

The goal of biometric identity verification is to assert whether a certain person is indeed whom he/she claims to be, based on behavioural or biological traits, also known as biometric modalities. Speech is a personal trait that can be used for biometric user identity

*Corresponding author. Tel.: +41 21 693 4691.
E-mail address: jonas.richiardi@epfl.ch (J. Richiardi).

0016-0032/$30.00
doi:10.1016/j.jfranklin.2006.07.002

verification and benefits from ease of use for the end user, cheap sensor hardware, a long research history, and can be deployed in a variety of environments.

Speaker verification performance is however very dependent on noise, in addition to intra-speaker variability which is strongly influenced by health factors, emotional state, and inter-session time. In the context of this paper, we define noise as "any unwanted change in a signal". Noise can be categorised as stationary, meaning its characteristics do not change with time, or non-stationary, meaning noise is itself a dynamic, time-dependent phenomenon. The influence of noise on clean speech can be modelled by the type of interaction the noise and speech signals have: additive noise is, as its name suggests, added to the speech signal. An example of additive stationary noise is fan noise from a desktop PC, and an example of additive non-stationary noise is a door slamming or mouth clicks (such as produced by the parting of the lips). Convolutional noise is due to the physical transmission chain of speech and can be caused by acoustical characteristics of the ambient environment, transducer operation, signal conditioning, signal processing, signal coding, and transmission channel. An example of stationary convolutional effect is caused by different frequency responses and self-noises characteristics of microphones, while non-stationary convolutional noise can be caused by dynamically changing reverberation conditions (due to room occupation change, pitch differences, etc.). In the present paper, we focus on stationary and non-stationary additive noise.

It is well known that the statistical distribution of common speech features such as MFCCs (Mel-Frequency Cepstral Coefficients) is significantly distorted when the original clean speech signal is subjected to additive noise [1,2] or convolutional noise [3]. Thus, the scores output from the classifier of a speaker verification system based on models of statistical distribution of features in clean acoustic conditions will notably change when presented with feature vectors corrupted by noise. In this situation some impostors will be able to obtain higher scores, and, respectively, some clients will obtain lower scores than in clean conditions, hence increasing false accept (FA) or false reject (FR) rates. To avoid such incorrect decisions the work presented in this paper aims at quantifying the amount of trust that should be put in a speaker verification classifier's decision, taking into account acoustic environmental conditions and behaviour of the classifier on evaluation data.

The problem of "knowing when the classifier is right", has seen numerous incarnations in very diverse areas and applications of pattern recognition, of which we will give but a few examples here. In handwriting recognition, Pitrelli and Perrone [4] have explored a number of confidence measures, the best of which encode a measure of the dispersion in scores of the top candidate words (in which case the confidence measures are for instance negative entropy, selectivity, or score ratios between candidates). In image orientation determination, Luo and Boutell [5] use various specialised object detectors (faces, grass, etc.) and low-level image features combined probabilistically to produce a confidence value which can be used to reject the detected image orientation. In protein localisation, Huang and Li [6] use a "reliability index" to estimate the certainty of the classification decision, which is computed based on the difference in score between the top and the second candidate class. In multimodal audio–video source localisation, Lo et al. [7] use cross-correlation of acoustic power per sector in a microphone array with reference profiles and amount of change in foreground-to-background ratio to derive reliability of audio localisation and motion detection. Finally, in speech recognition, confidence measures have seen wide usage in the past 10 years, and a large number of methods have been

proposed incorporating evidence from the acoustic, grammatical, pragmatical and classifier output domains [8].

In the remainder of this paper, we focus the discussion on confidence measures in speaker verification (Section 2), and we divide approaches into two groups depending on the domain of the evidence they are based on (classifier-domain only or multiple domains). We then explain in more details the operation of a recently proposed approach belonging to the second group, called reliability estimation (Section 3), before giving examples of how the confidence and reliability measures can be used (Section 4). We present experimental results for confidence and reliability approaches on a noisy database (Section 5), and finish with concluding remarks (Section 6).

2. Confidence measures in speaker verification

In speaker verification, confidence measures have been used for various applications and purposes. In forensic cases, they have been proposed to indicate the degree of belief the court should place in a classifier's output score [9,10]. In on-line speaker model adaptation, it has been suggested to use them to perform unsupervised selection of adaptation data [11]. In multimodal verification, confidence measures are increasingly used to provide weights for the fusion algorithm [12]. In text-independent speaker recognition tasks, some authors have proposed their use to defer the decision to manual or automated post-processing [13].

Regardless of the application, we distinguish two broad types of confidence measures in speaker verification: those that use only data provided by the classifier, such as likelihood scores, likelihood ratio scores, or hard decisions, and those that use auxiliary information from multiple domains, for instance signal-domain quantities such as fundamental frequency or signal-to-noise ratio (SNR) combined with classifier-domain information.

2.1. Confidence measures with classifier-domain data

In the following discussion we strive to present approaches keeping a consistent notation throughout. To this end, we first define variables of interest for speaker identity verification. *TID* (True user IDentity), corresponds to the ground truth, and *CID* (Classified user IDentity) corresponds to the speaker verification classifier's decision. $TID = 1$ represents the event "the system user is a client", while $TID = 0$ corresponds to the event "the system user is an impostor". $CID = \{0, 1\}$ corresponds to the events "the classifier accepts the identity claim" ($CID = 1$) and "the classifier rejects the identity claim" ($CID = 0$). As a shorthand, we introduce another binary variable *DR* corresponding to "decision reliability", where $DR = 1$ represents the statement "the classifier is correct" and $DR = 0$ represents the opposite statement. In Boolean logic terms, $DR = \overline{CID \oplus TID}$. A fundamental quantity in speaker verification is the log-likelihood ratio score, which we denote *Sc*. It represents the log of the ratio of the likelihood of the utterance (biometric presentation) given a particular client model to the likelihood of that presentation given a background model.

The distribution of verification scores *Sc* can serve as a basis for simple confidence measures. Nakasone and Beck [14] propose a Bayesian confidence measure which can be expressed in speaker verification terms as the posterior probability that the utterance is

from a client given the score:

$$P(TID = 1|Sc) = \frac{P(TID = 1)P(Sc|TID = 1)}{\sum_{id=0}^{id=1} P(TID = id)P(Sc|TID = id)}, \quad (1)$$

where *id* represents either an impostor ($id = 0$) or a client ($id = 1$). By assuming that the client and impostor score distributions are Gaussian (which is often not true), they then define the confidence measure by fitting a logistic function to the posterior probability represented by Eq. (1):

$$CM_{\text{logistic}}(Sc) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Sc)}}, \quad (2)$$

where in our implementation the β exponential parameters are learned using a least-squares method. We adapt this measure from the forensic context by also computing $P(TID = 0|Sc)$ with a change of numerator in Eq. (1), then fitting a decreasing sigmoid $1 - CM_{\text{logistic}}$ to that posterior. This allows us to use this measure for the negative identification case also. One further change is needed since the ground truth is not available during testing. Thus, we replace *TID* with the classifier's opinion *CID*, and use the appropriate measure at runtime depending on the classification result.

This measure presents two main drawbacks: it assumes Gaussian class-conditional distributions for scores, and does not take into account the actual error distributions of the classifier. These two drawbacks can also be seen as strong structural constraints that prevent overfitting and mean that this model may generalise better given a small amount of training data.

In speaker identification, Gish and Schmidt [13] rely on the reasonable assumption that the scores of the top candidates in the case of correct classification is higher than those of incorrectly identified candidates. Their modelling is based on two distributions: The distribution of scores for incorrect classifications $P(Sc|DR = 0)$ (hereafter abbreviated $P_{\text{wc}}(Sc)$) and correct classifications $P(Sc|DR = 1)$[1] (hereafter abbreviated $P_{\text{cc}}(Sc)$). This is an interesting approach, since most confidence measures in speaker verification, and indeed in other fields of pattern recognition, are centred on the class-conditional distributions of scores. They propose two methods to evaluate confidence in speaker identification applications, one based on significance testing, and the other on a Bayesian posterior probability $P(DR = 1|Sc)$.

The confidence measure based on significance testing is expressed thus:

$$CM_{\text{sig}}(Sc) = 1 - \int_{Sc}^{\infty} P_{\text{wc}}(Sc)dSc. \quad (3)$$

However, this significance-based confidence measure cannot be readily adapted to the verification case, because it essentially measures "how far on the tail of the distribution of incorrect classification scores the observed score occurs" (see Fig. 1(a)), which is appropriate for identification, but not for verification. Indeed, while it can be expected and assumed that the mean of the $P_{\text{wc}}(Sc)$ distribution will be lower than the mean of the $P_{\text{cc}}(Sc)$ distribution, in verification the errors will be clustered around the threshold and

[1]It should be noted that in the identification context *Sc* is an identification score, not a log-likelihood ratio as used in verification. Also, the semantics of *DR* change to $DR = 1$ if the candidate corresponding to the top identification score is indeed the target, and $DR = 0$ otherwise.

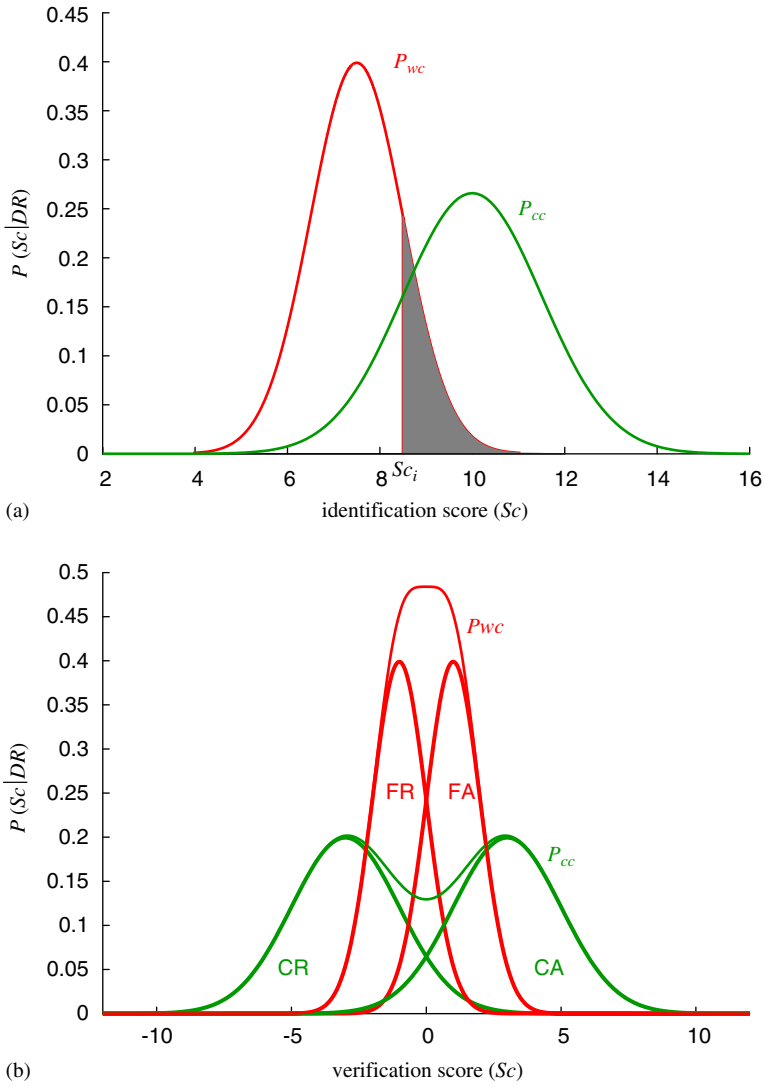

Fig. 1. Idealised score distributions for classifier errors and correct decisions in identification and verification: (a) idealised score distributions for correct identification ($P_{cc}(Sc)$) and identification error ($P_{wc}(Sc)$). The solid grey area under the distribution of identification error scores corresponds to the second term in Eq. (3), where a particular identification score Sc_i determines the lower bound of integration; (b) idealised graph of correct verification ($P_{cc}(Sc)$) and verification error ($P_{wc}(Sc)$) score distributions showing the four sub-distributions: correct reject (CR), false reject (FR), false accept (FA), and correct accept (CA). Note that in reality the sub-distributions are likely to be non-Gaussian and overlap in a different way.

correct decisions can be taken both below the threshold (correct reject) and above the threshold (correct accept). Thus, the confidence measure would need to be symmetric around the threshold, to account for false reject errors. A second important observation is that a confidence measure based on the $P_{cc}(Sc)$ and $P_{wc}(Sc)$ distributions in verification

needs to take into account the bimodal nature of the correct decision score distributions. This point is illustrated in Fig. 1(b).

Gish and Schmidt also propose a Bayesian confidence measure, which quantifies the posterior probability that the identification decision is correct given the score:

$$CM_{\text{Bayes}}(Sc) = P(DR = 1|Sc) = \frac{p_{\text{cc}} P_{\text{cc}}(Sc)}{p_{\text{cc}} P_{\text{cc}}(Sc) + p_{\text{wc}} P_{\text{wc}}(Sc)}, \tag{4}$$

where p_{cc} is the prior probability that the classification is correct, and p_{wc} is the prior probability that the classification is wrong. In identification, this can be estimated from results on an evaluation set. This measure can be applied in verification, but to set the priors an operating point must be chosen which corresponds to a particular threshold setting. An example for this is to choose the percentage of errors on an evaluation set, setting $p_{\text{wc}} = N(DR = 0)/N, p_{\text{cc}} = 1 - p_{\text{wc}}$ (where N is the total number of test cases in the evaluation set) ensures proper normalisation. For a well-performing speaker verification system, the ratio $p_{\text{cc}}/p_{\text{wc}}$ is 10 or more. Thus, the confidence measure will be biased high and will most likely report high confidence. This can be compensated by using non-informative priors, meaning the confidence measure will be based only on the score distributions, without taking into account the priors. If the $P_{\text{cc}}(Sc)$ and $P_{\text{wc}}(Sc)$ score distributions were modelled as mixture distributions, this confidence measure should provide good accuracy when applied to verification tasks given that the score distributions are trained on an evaluation set which comes from an environment acoustically similar to that of the test set. However, in this paper we keep with the original definition of the measure for speaker identification and model scores using one normal distribution for each of $P_{\text{cc}}(Sc)$ and $P_{\text{wc}}(Sc)$.

Poh and Bengio [12] use the FR rate for a certain score (taken as threshold) subtracted from the FA rate for the same threshold. Thus, the closer the score is to the decision threshold, the lower the confidence:

$$CM_{\text{margin}}(Sc) = |FAR(Sc) - FRR(Sc)|. \tag{5}$$

The client and impostor distributions are trained on an evaluation set. To avoid condition mismatch leading to erroneous test results, these functions can be trained on an evaluation set using conditions similar to those present in test conditions. This approach is interesting because it takes into account the distribution of errors with respect to a score, and it is quite generic: the sources of noise (both additive and convolutional) are subsumed and abstracted by their effects on the score distributions. This is far less complex than trying to model noise and distortions in the signal domain. The authors then show a theoretical framework for combining this confidence measure with a speech quality measure (QM) in order to enhance fusion in multimodal biometrics.

As hinted by the notation used above, the confidence measures can be used as functions of the verification score. Thus, to obtain a better intuitive understanding of the various confidence measures presented, it is interesting to plot a graph of the output of the function with respect to the input score. This is depicted in Fig. 2. We also introduce the posterior of the reliability approach which will be presented in Section 3. Since it takes into account the amount of noise in the signal, this posterior will change dynamically with each presentation, with the concave part changing shape with increasing levels of noise.

A few other methods have been used to estimate confidence in speaker verification or identification (notably [15,16,11]), but not all of these are bounded to the [0; 1] interval and

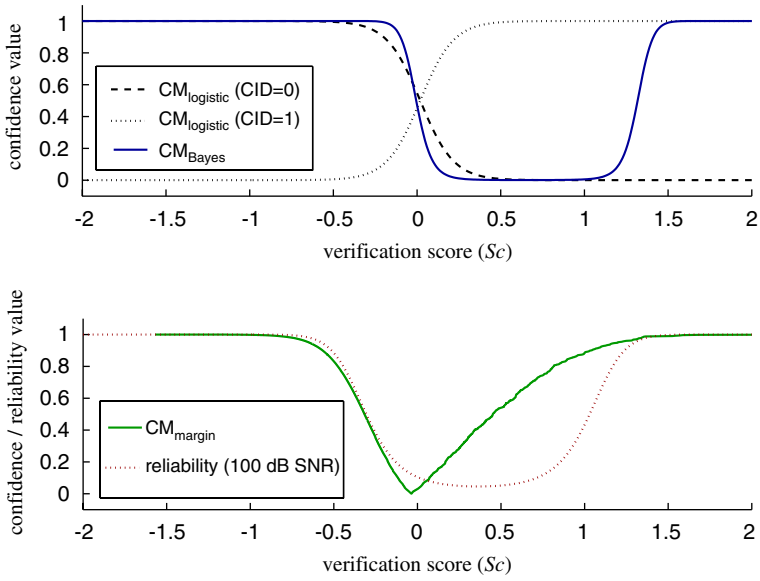

Fig. 2. Output of confidence and reliability measures with respect to presentation score. The reliability curve is for an artificially fixed signal-to-noise ratio of 100 dB.

would need to be transformed (for instance using sigmoid mapping) to facilitate comparison with the other methods.

This brief presentation of classifier-domain confidence measures in speaker verification and identification reviewed only text-independent confidence measures; in passing we should point out that many confidence measures defined in speech recognition can also be applied to text-dependent speaker verification (see for instance [17]).

2.2. *Confidence measures with data from multiple domains*

Some authors have also seen the need to incorporate other sources of information than just the classifier's output into their confidence estimate. For instance, it is widely recognised that shorter test utterances give less reliable results. Likewise, noise in the signal domain contributes to augmenting the error rate of speaker verification systems.

Recently, Campbell et al. [10] have used signal-domain data (utterance duration, channel-type label and SNR) in addition to utterance score to estimate a confidence for each score. Since the application is in forensics, the task is to compare whether two utterances come from the same speaker, using a speaker verification system. Thus, this confidence measure can be expressed as "the a posteriori probability that the two utterances come from the same speaker". The confidence model is a multi-layer perceptron (MLP), bringing the benefit that the shape of the distributions does not have to be assumed a priori.

In speaker identification, Huggins and Grieco [18] also take into account additional information beyond signal-domain quantities, such as amount of overlap between models in feature space. Their main indication of confidence is a combination of training/testing utterance duration and SNR. Their method is based on computing error rates with respect

to seven discrete SNR levels (pink noise mixed in from 6 to 24 dB) for 13 different utterance durations, thus resulting in 91 regression models indexed by training and testing utterance duration. The information about the amount of overlap between models can then be factored in by computing another regression on top of the basic duration/SNR combination, which then results in a prediction error of $4.2 \pm 3.3\%$ on a 40-speakers database. One issue that is reported is that the model may be overly relying on SNR as its main indication of classifier performance, as the accuracy of confidence prediction drops when training and testing environments are matched. It is not clear how additional measures of quality would be added to the model. The confidence measure derived does not lend itself easily to a probabilistic interpretation.

Richiardi et al. [19,20] have proposed a measure of confidence taking into account signal-domain QMs in the form of different estimations of the SNR. Their confidence is estimated by a graphical model trained on score distributions produced during erroneous and correct decisions, and the signal QMs just mentioned. We will now present more details about this last method.

3. Classifier decision reliability in speaker verification

The approach we propose[2] estimates a quantity, the *reliability of the classifier's decision*, which can be phrased as "the probability of having taken a correct classification decision given available evidence". The evidence we use and the modelling framework differ from the "confidence measure" approaches we reviewed in Section 2 in several ways, some of which we will expand on here. We define the estimation of reliability as an interpretable *probabilistic* method providing an output in the form of a posterior probability, based on combining *classifier error modelling* and *signal-domain* information.

Our approach combines information about both the acoustic environment and the classifier behaviour in order to provide decision reliability information in speaker verification. In this context, the result of speaker verification is directly influenced by the real state of the user identity (client or impostor) and the state of the decision reliability measure, given additional evidence about the environmental acoustic conditions. Since the intercausal relations of these two factors cannot be established deterministically, we use a probabilistic reliability measure. Instead of being assigned a particular value, a probabilistic reliability measure is defined by a distribution over its possible values. In our approach we use Bayesian networks (BNs) for inferring the decision reliability distribution.

3.1. Bayesian networks for reliability estimation

BNs, also called belief networks, are graphical models used to describe a joint probability distribution (pdf) over a finite set of random variables [21], and are completely defined by the triple (V, A, C), where V is the set of nodes associated with the random variables, A is the set of arcs and C is the set of conditional probability distributions associated with the nodes' variables. The arcs between the nodes point from all parent variables to their children variables. The intuition behind directionality represents the fact that the parent variables can directly influence their children and this influence can be

[2]This paper is an extension to the work presented at ICASSP 2005 [19].

interpreted as a cause-effect relationship. The joint pdf represented by the BN can be written as a product of all nodes' CPDs (conditional probability densities of each node given its parents). Finally, the basic task for any BN is to perform inference, that is to compute the posterior distribution for a set of "query" variables, given some observed event, i.e. evidence for some observed variables.

The BN in Fig. 3(a) depicts an influence model for the variables *TID*, *CID* and *DR*. In this network, the fact that the user's identity claim is true or not can be seen as the cause of a particular classified user identity value, unless the classifier performs at random. That is, for a working speaker verification system it is more likely that a client attempt results in $CID = 1$ than $CID = 0$. The decision reliability can be seen as an alternative cause that might also point at errors in the *CID* value. For example $CID = 1$ can be explained by $TID = 1$ and $DR = 1$ (the classifier makes a correct verification decision because the user is a client and the decision is reliable) or $TID = 0$ and $DR = 0$ (the classifier makes a wrong verification decision because the user is an impostor and the decision is unreliable).

In this case, having the set of nodes associated with the random variables $V = (TID, CID, DR)$, and taking into account the arcs defined in Fig. 3(a), the joint pdf over V can be written as

$$P(TID, CID, DR) = P(DR)P(CID|DR, TID)P(TID). \tag{6}$$

Since the variables *TID* and *DR* are not observable during speaker verification, we need to provide additional sources of information that can be observed and can provide evidence in favour of particular (TID, DR) values. The verification score (likelihood ratio, see Section 5.2) is known to carry information about the state of the user identity (client/impostor), while a signal QM can be used to provide evidence for the *DR* variable. For example, the SNR of the speech signal can be used to measure the level of the acoustic noise. Therefore, we define the two continuous variables *Sc* and *QM*, corresponding to the score of the verification and the QM for the given modality (SNR in the case of speech). *DR*, *CID* and *TID* can be seen as causes for the observed *Sc* value, while *DR* can be seen as the cause for *QM* values.

The final version of the BN incorporating all these variables $V = (TID, CID, DR, Sc, QM)$ is depicted in Fig. 3(b). Discrete variables are drawn as squares, and circles are used for the continuous ones. The CPD of discrete variables is represented by a probability table, while continuous variables make use of arbitrary parametric CPDs. We use conditional Gaussian distributions.

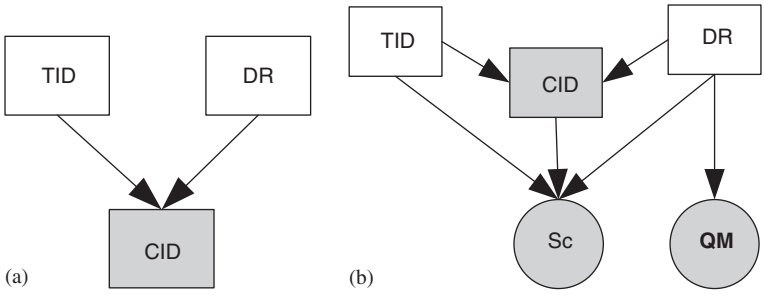

Fig. 3. Bayesian networks for decision reliability estimation: (a) reliability model with classifier-domain evidence; (b) full reliability model with score-domain and signal-domain evidence.

In our case the posterior

$$Rel(Sc, CID, QM) = P(DR|CID, Sc, QM) \tag{7}$$

is the distribution of decision reliability measure. To mark an observed variable in the graphical representation of the BN we use shading, and unobserved variables are left white. Once the CPD functions for all the nodes given their parents are defined, exact or approximate inference on each node in the network can be performed. Since the number of variables in our case is small, exact inference on a junction tree algorithm can be applied [21].

A similar architecture has been used by Toyama and Horvitz [22] for a head tracking application in computer vision. In their case, several visual tracking algorithms are combined and the reliability of each is estimated before taking a final decision. One difference with our approach is that the domain of QM we use is directly related to the signal, and is not dependent on the classification algorithm, whereas they identify "failure modes" for their tracking algorithms and use these specificities as features. A second difference is that we explicitly define and label DR as a binary variable during training with a fixed semantic meaning, rather than keeping the value for this node hidden also during training. Lastly, our topology has additional arcs not found in [22].

Before the BN can be used to produce reliability estimates, the conditional distributions defined by its topology must be learned.

3.2. Training procedure for Bayesian network reliability estimator

In order to perform inference on $P(DR|CID, Sc, QM)$ the conditional probability distribution parameters for the network variables have to be learned from training examples. In the case of fully observable variables in the training set, the estimation can be done with random initialization and maximum likelihood (ML) training [21]. After the training, the posterior distribution $P(DR|CID, Sc, QM)$ can be used as a probabilistic measure of speaker verification decision reliability.

The training setup for the BN is depicted in Fig. 4. A speaker verification system provides values for the CID, Sc variables, and an acoustic environmental condition measure provides the QM values. We use an SNR-related measure described in Section 3.3. To simulate the effects of a degraded acoustic environment, babble-type noise corresponding to a possible deployment environment with SNRs following a random uniform distribution from 5 to 55 dB is added to the database. This noisy speech data is an input source for the verification system, which calculates Sc and sets CID according to the threshold for that user. According to the match between the true identity of the speaker (TID) and the speaker verifier output (CID), we label DR in the BN training data as being "true" ($TID = CID$ then $DR = 1$) or "false" ($TID \neq CID$ then $DR = 0$). In that way, we model the relationship between verification errors and environmental conditions. We assume the speaker verification classifier performs above chance level in clean conditions.

Since we also assume the speaker verification classifier performs above chance in noisy conditions, the data set used for BN training will by definition always contain less data labelled $DR = 0$ ($TID \neq CID$) than data labelled $DR = 1$. An additional source of imbalance is that for most biometric databases, the size of client data ($TID = 1$) is smaller than impostor data ($TID = 0$); with the random impostors technique ("pseudo-impostors") any other user can serve as an impostor to a particular user.

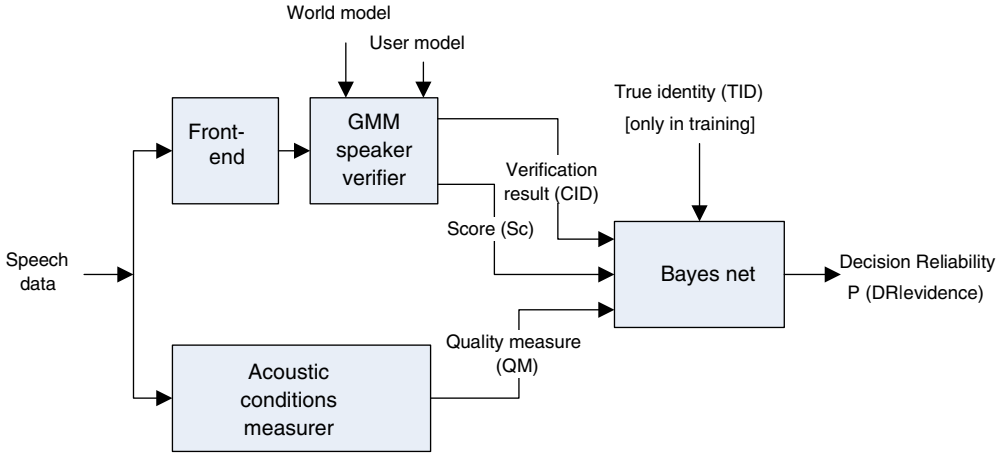

Fig. 4. Combined speaker verification and decision reliability estimation systems.

During training of the BN, the prior probabilities over the *DR* and *TID* variables are assigned according to the counts of data samples corresponding to the different *DR* or *TID* variable realisations. Thus, if the BN is trained without special care, the learned prior probabilities on the *DR* and *TID* variables will be mismatched for client and impostor access because the counts for impostor accesses will be much higher. The training of a reliability measure suffers from a double imbalance problem, since both the final class of interest (*DR*) and an important factor in system (impostor or client access) are by nature imbalanced. Since we do not want to bias the final posteriors over the *DR* values on the number of data counts, it is important to balance the number of examples for client and impostor access (*TID* variable). In addition, because *DR* is a competing cause for the explanation of *CID* it is also important to balance the number of examples with respect to the *DR* values.

Three approaches exist to handle the imbalanced sample problem: undersample the class with the largest number of training examples, oversample the class with the smallest number of training samples, or bias the learning procedure [23]. In our method, we use the latter by fixing both the $P(TID)$ and $P(DR)$ priors to be uniform (0.5). The BN formulation offers an easy way to accomplish this, by simply ''clamping'' the nodes whose probability distributions we do not wish to learn from data.

3.3. Taking acoustic conditions into account

To measure the acoustic conditions for the speech we use a SNR-related measure. The SNR can be defined as the ratio of the average energy of the speech signal divided by the average energy of the acoustic noise in dB. As we use a single-channel speech signal we estimate these energies based on a voice activity detection (VAD) and corresponding speech/pause segmentation. The VAD algorithm is based on the ''Murphy algorithm'' described in [24]. We then assume that the average energy of pauses is associated with that of noise. An SNR-related signal QM is given by the formula:

$$QM = 10\log_{10}\frac{\sum_{i=1}^{N} Is(i)s^2(i)}{\sum_{i=1}^{N} In(i)s^2(i)}, \tag{8}$$

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)

NAS_μ (%)	SAN_μ (%)	R_μ (%)
13.03	11.45	12.47

All results are averaged over the utterances in the *individuals* set of the CUAVE database.

where $\{s(i)\}, i = 1, \ldots, N$ is the acquired speech signal containing N samples, $Is(i)$ and $In(i)$ are the indicator functions of the current sample $s(i)$ being speech or noise during pauses (e.g. $Is(i) = 1$ if $s(i)$ is a speech sample, $Is(i) = 0$ otherwise) as reported by the voice activity detector.

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 [25]. This is a labelled database containing 36 individual users, both male and female, each providing utterances of separated digits for about 2 min. The performance is computed in terms of four quantities [26]: *front-end clipping* (*FEC*), indicating speech misclassified 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 three 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 misclassified, 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 SNR. Therefore, the VAD will be less accurate on noisy data.

The BN defined is flexible enough to accommodate several speech signal QMs. Indeed, the estimate used is based on energy and may not provide consistent speech/pause segmentation boundaries in very noisy conditions, thus defeating the purpose of estimating the SNR in the first place. An interesting extension is to use more robust segmentation algorithms, such as those based on entropy [27], to augment the QM scalar and turn it into a vector of QMs. This simply makes the corresponding Gaussian *QM* node multivariate instead of univariate [20].

4. Using reliability and confidence measures in speaker verification

4.1. Classification with the reject option

Many decision errors in biometric verification are due to ergonomic factors rather than algorithmic weaknesses. For instance, in iris and face verification improper distance and centering of the image can significantly degrade verification accuracy. In speech-based verification, distance from the microphone and speaking volume are important factors. In this section, we propose that the best strategy to cope with uncertain classification results is to re-acquire the signal up to N times rather than try to compensate signal-domain noise by

other means. This idea is present in other fields of pattern recognition such as optical character recognition, where an example is that the second-stage recogniser can reject the character segmentation proposed by the preprocessing module if the confidence value associated to it is too low [28].

In an interactive speaker verification system, the user could be asked to move closer to the microphone if the SNR is too low, or the operator could be informed that verification results for presentation n are unreliable. In this case, performing sequential repair only if needed presents the advantage of minimising the amount of interaction between the user and the system, thus speeding up the verification process. The final classifier decision *FCID* can then be presented as a definitive verification result.

The sequential repair strategy outlined in Fig. 5 is equivalent to doing intra-modal fusion with binary weights at the score level, where the score of the unreliable presentation(s) gets a weight of 0 and the reliable presentation gets a weight of 1. Instead of throwing away all the information provided by the first presentation, it is possible to combine it with the second presentation. A simple scheme is to weight each presentation score by its corresponding normalised reliability value to derive the final (fused) score:

$$Sc = \sum_n Rel(Sc_n) \cdot Sc_n, \tag{9}$$

where the normalised reliability values are obtained in the following fashion:

$$Rel(Sc_n) = \frac{P(DR = 1|CID_n, Sc_n, QM_n)}{\sum_n P(DR = 1|CID_n, Sc_n, QM_n)}. \tag{10}$$

In this case, the decision to acquire a new presentation would still be governed by the insufficient reliability of the first presentation. The advantage of this scheme over a scheme that would always acquire two presentations is that the interaction time with the speaker verification system can be minimised. By setting the reliability threshold, it is possible to bias the system towards being more tolerant of low reliabilities (resulting in higher error rates), or less tolerant (resulting in longer interaction time with the system for users).

4.2. Interpretation of strength of evidence in forensic applications

Another use for the reliability measure is as a tool to help the forensic expert quantify the degree of trust that she should put in the opinion given by an automatic system.

The main goal of forensic speaker recognition is to interpret evidence material in the course of a criminal investigation. In the case of questioned recording (trace), the evidence does not consist in speech itself, but in the quantified degree of similarity between speaker dependent features extracted from the trace, and speaker dependent features extracted from recorded speech of a suspect, represented by his/her model. In an automatic approach, this similarity measure is quantified by a similarity score (e.g. log-likelihood if the suspected speaker is represented by Gaussian mixture model (GMM)).

The calculated value of evidence does not allow the forensic expert alone to make an inference on the identity of the speaker. It can be done using the strength of evidence, expressed in terms of the likelihood ratio of the evidence given two competing hypotheses: (1) the suspected speaker is the source of the questioned recording ($TID = id$), (2) the speaker at the origin of the questioned recording is not the suspected speaker ($TID \neq id$).

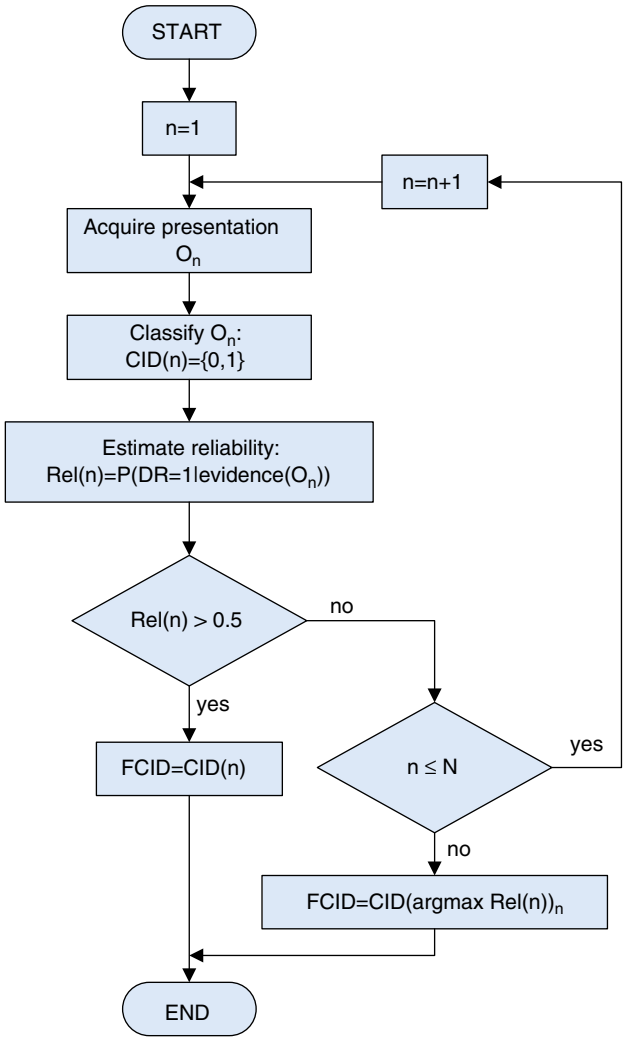

Fig. 5. Sequential repair sequence for reject option.

The likelihood ratio calculated in the odds form of Bayes' theorem (LR in Eq. (11) below) summarizes the statement of the forensic expert in the casework [29].

The odds form of Bayes' theorem in Eq. (11) shows how new data (evidence) can be combined with prior background knowledge (prior odds (province of the court)) to give posterior odds (province of the court) for judicial outcomes. The prior odds cannot be set by the forensic expert because she has no access to the whole prior background knowledge.

$$\frac{P(TID = id|E)}{1 - P(TID = id|E)} = \frac{P(TID = id)}{1 - P(TID = id)} \times \underbrace{\frac{P(E|TID = id)}{P(E|TID \neq id)}}_{LR}. \tag{11}$$

Consequently, the greatest interest to the court is the extent to which the likelihood ratios correctly discriminate "same speaker and different-speaker" pairs under operating conditions similar to those as regards to the case in hand. If the prior odds can be estimated by the court, evaluating confidence and reliability in the scores domain can become a key aspect of the speaker recognition problem before final decision of the court. Because reliability explicitly models the effect of acoustic conditions on the score distributions, the common problem of operating conditions mismatch can be alleviated.

On the other hand, the performance and reliability of an automatic speaker recognition method should be evaluated and presented to the court by the forensic expert. This can be done by repeating the experiment of likelihood ratio calculation, with several speakers being at the origin of the questioned recording, and by representing the results using two probability distributions of likelihood ratios when the two competing hypotheses are true [30]. In this case, the method based on reliability measures, presented in this paper, can be used after the replacement of scores by likelihood ratios given a threshold of likelihood ratio equal to one.

5. Experiments and results

5.1. Assessment criteria

The first measure of performance that we use for assessing confidence and reliability measures is the accuracy of prediction of decision correctness. As mentioned in Section 3.2, the number of samples per class[3](clients, $TID = 1$, and impostors, $TID = 0$) is imbalanced (around 1:250 in our case), hence we cannot take the classical definition of accuracy as $nCorrectClassifications/nSamples$, or the performance of the confidence and reliability measures for client accesses would have very little influence on the overall results. Furthermore, since the baseline classifier will have an error rate of less than 50% (otherwise it should not be used), there will always be less cases where $DR = 0$ than cases where $DR = 1$. Thus, a blind confidence measure could predict $DR = 0$ all the time and be mostly correct if this imbalance is not accounted for. Since we have a "double imbalance" situation, we do not make use of the geometric mean which can be useful in "single imbalance" situations [23,31], but rather we define balanced accuracy as

$$acc_{\text{bal}} = \frac{1}{4} \sum_{dr=\{0,1\}} \sum_{tid=\{0,1\}} \frac{N\text{corr}_{DR=dr,TID=tid}}{N_{DR=dr,TID=tid}}, \tag{12}$$

where $N\text{corr}_{DR=dr,TID=tid}$ is the number of correctly classified samples out of a total of $N_{DR=dr,TID=tid}$ samples with ground truth labels $DR = dr$ and $TID = tid$. This measure expresses the overall performance of the reliability or confidence measure. A measure that performs well for, say, impostors, but not for clients will thus be penalised by this evaluation criterion.

The performance of confidence measures over a set of test data can also be evaluated by producing a detection error tradeoff (DET) curve based on two distributions of confidence or reliability measures: one for the measures over correct decisions ($DR = 1$), and one for

[3]In the following discussions, *class* will mean impostor ($TID = 0$) or client ($TID = 1$) access when talking about the speaker verification classifier. When talking about the reliability or confidence measure, which can be considered as a second-level classifier, *class* will be taken to mean correct ($DR = 1$) or incorrect ($DR = 0$) decision.

the measures over wrong decisions ($DR = 0$). The less overlap between the distributions there is, the better the confidence or reliability measure will be. DET curves are a meaningful tool to compare confidence and reliability measures only if these are trained with the same assumptions about the imbalance of the training set. In the present case, CM_{logistic} (with uniform priors on *TID*), CM_{margin} (with equal cost for FA and FR in building the FAR, FRR curves) and reliability (with uniform priors on *TID* and *DR*) can be compared, because the structure of the testing set in terms of *TID–DR* class balance will have little impact on the results of the test. CM_{Bayes} however is based on direct modelling of the correct and erroneous decisions score distributions (CA, CR, FA, and FR, see Fig. 1(b)) and thus will be favoured by a test set structure matching the training set structure (small data counts for CA, FR with respect to CR, FA). Therefore, direct comparison makes little sense.

Another objective measure of goodness for reliability or confidence measures is normalised cross-entropy (normalised mutual information). It can be defined as the "relative decrease in uncertainty about the classifier's decision provided by the confidence measure", while the original definition from NIST for speech recognition confidence measures [32] is "the mutual information (cross entropy) between the correctness of the system's output word and the confidence score output for it, normalized by maximum cross entropy". However, this measure is also biased in favour of confidence or reliability estimates that perform better on the majority class ($DR = 1$). Thus, while it is very useful in speech recognition applications, we do not use it for evaluation in the current biometric identity verification setting given the imbalance of classes.

5.2. Speaker verification system

The speaker verification system, based on the Alize Toolkit [33], uses energy-based VAD to remove pause portions of the utterance, after which 12 MFCCs with first- and second-order time derivatives are extracted using the HTK toolkit's `HCopy`, and cepstral mean normalisation is applied to try and compensate for stationary convolutional noise. All training files for the users in the database are pooled to train a world GMM with 200 Gaussian components with diagonal covariance matrices. Each user's model is then MAP-adapted from the world model by using all of the corresponding training files.

A global verification threshold is obtained by using all utterances from the evaluation set to obtain verification scores, and setting the verification threshold at the EER (equal error rate) point between client and impostor score distributions.

5.3. Database and results

The database used for experiments is a 251-users subset of TIMIT, divided in 179 males and 72 females. Approximately, 20 s of data in six presentations (phonetically rich read sentences) for each user is used to train the user models. The remaining four client presentations per user are held out and divided into evaluation (two presentations) and testing (two presentations) sets. In verification, no separation is made between male and female pool.

We also use a noisy version of the same TIMIT subset: babble-type noise corresponding to a possible deployment environment with SNRs following a random uniform distribution from 5 to 55 dB is added to the clean data.

5.3.1. Decision correctness prediction

The aim of this series of experiments is to empirically verify how well the confidence and reliability measures perform on a held-out testing set. We treat confidence and reliability estimation as a two-class pattern recognition task, where the goal is to infer the state of the DR variable. The measures are trained on evaluation data, which can be either noisy (thus matching the noisy test environment better) or clean.

All these measures are trained with equal priors whenever possible to make the balanced accuracy comparison meaningful. For instance, CM_{logistic} is trained with uniform priors on TID, CM_{Bayes} with equal priors for p_{cc} and p_{wc}, CM_{margin} with equal cost for FA and FR when computing the FAR and FRR curves, and reliability with equal priors on TID and DR.

All measures tested perform better than chance (25% on a four-class problem). Thus, their use should prove beneficial on a testing set that is evenly balanced between CA, CR, FA, and FR cases, because erroneous and correct decisions from the speaker verification system will be identified correctly. Secondly, it should be noted that measures that do not make Gaussian assumptions about the impostor/client score distributions (CM_{margin} and reliability) perform overall better than the others. The slightly better result obtained by reliability can be explained by the fact that signal quality is taken into account (reliability is the only multiple-domain measure tested in this paper). The results of CM_{Bayes} could probably be made better with respect to the balanced accuracy criterion by substituting mixture models for P_{cc} and P_{wc}, where the weight given to client and impostor presentations would be equal.

From the results in Table 2, it can be seen that the three confidence measures tested perform better or only slightly worse when trained on noisy data. This is in accordance with classical results for simple robustness methods in speech recognition.

To appreciate the behaviour of the confidence and reliability measures over a range of thresholds for the $DR = \{0, 1\}$ decision, the performance of the confidence and reliability measures trained on the noisy evaluation set and tested over the noisy testing set is shown in Fig. 6. This figure also displays the results for reliability when the priors on $P(TID)$ and $P(DR)$ are not fixed to be uniform but learned from training data. As can be expected, since the testing data matches the training data in terms of respective class counts, the result is better than for the reliability measure trained with uniform priors.

Table 2
Decision correctness prediction for reliability and confidence measures. acc_{CA} is the accuracy on correct accept cases ($TID = 1, DR = 1$), acc_{CR} is the accuracy on correct reject cases ($TID = 0, DR = 1$), acc_{FA} is the accuracy on false accept cases ($TID = 0, DR = 0$), acc_{FR} is the accuracy on false reject cases ($TID = 1, DR = 0$), and acc_{bal} is the balanced accuracy computed as per Eq. (12)

Method	acc_{CA} (%)	acc_{CR} (%)	acc_{FA} (%)	acc_{FR} (%)	acc_{bal} (%)
CM_{logistic} (clean)	100	99.7	4.4	0.0	51.0
CM_{logistic} (noisy)	100	96.7	30.0	0.0	56.7
CM_{Bayes} (clean)	8.3	96.1	33.3	100	59.4
CM_{Bayes} (noisy)	5.8	95.4	35.6	100	59.2
CM_{margin} (clean)	43.9	49.4	91.1	98.8	70.8
CM_{margin} (noisy)	66.8	49.1	91.1	92.9	75.0
Reliability (noisy)	66.5	72.7	76.7	93.3	77.3

All accuracies are given in percent.

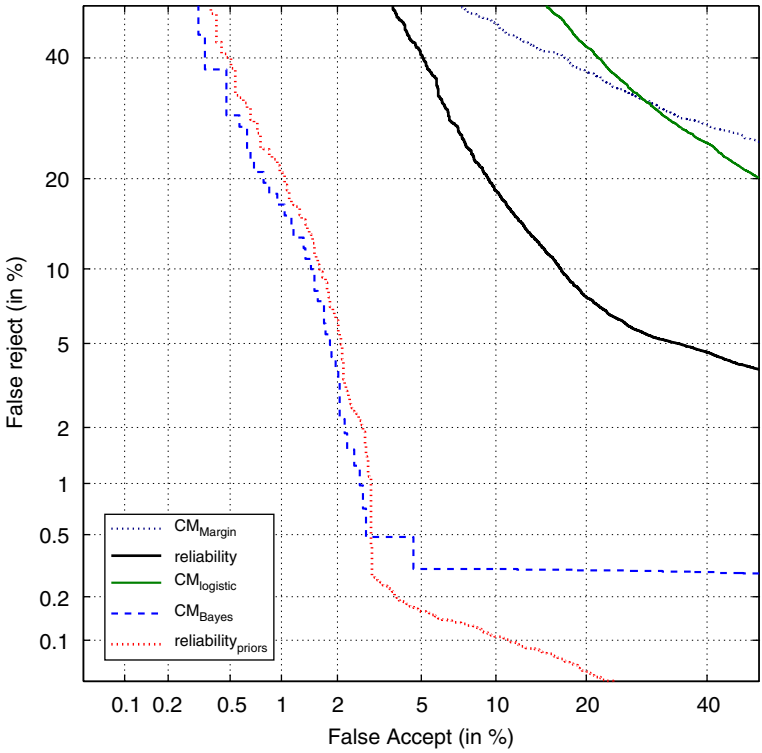

Fig. 6. DET curves for confidence measures and reliability measure, trained on noisy evaluation data and tested on noisy test data. *reliability*$_{\text{priors}}$ is trained with $P(TID)$ and $P(DR)$ learned from data, and thus results are better than for the reliability with non-informative priors since the testing set structure closely matches the training set structure.

This flexibility in setting the priors is an advantage of the reliability approach with BNs: the priors $P(TID)$ and $P(DR)$ can be instantly modified to suit the deployment conditions and scenario without having to retrain the rest of the network (conditional mixture of Gaussians for Sc and QM).

The relatively good performance for the CM_{Bayes} measure shown in Fig. 6 can again be attributed to the design of the test set, whose proportion of client and impostors (corresponding to correct accept and correct rejects, FAs and FRs) matches exactly that which is found in training. Given the small amount of correct accepts with respect to the number of correct rejects, and likewise the small amount of false rejects with respect to FAs, the $P_{\text{cc}}(Sc)$ and $P_{\text{wc}}(Sc)$ densities estimated and used for Eq. (4) (the CM_{Bayes} confidence measure) are strongly biased towards the majority class (impostors). Therefore, the matching testing set used favours this measure. It can be seen from the results in Table 2 that if the testing imbalance is removed by the evaluation criterion, the actual error rate is higher.

5.3.2. *Improving speaker verification performance*

To assess the potential of the reliability measure in improving speaker verification performance, we compare the performance of two repair strategies for the same speaker

verification system. The first repair strategy implements the simple repair sequence described in Fig. 5: the system is allowed to request another presentation if it estimates that the decision is unreliable (i.e. $P(DR = 1|CID, QM, Sc) < Threshold$) for the first presentation. If the second presentation also has low reliability, the system picks the presentation with highest reliability to produce the final score and accept/reject decision. In this case the behaviour is similar to a *max* rule on *DR* for intra-modality fusion. The results for this strategy, where the threshold has been set so that about one out of two test cases are re-acquired, are shown in Fig. 7 under the label "conditional most reliable of two". We can see that, at EER, this achieves the same error rate as a baseline method which randomly re-acquires presentation for 50% of the test samples and then takes the mean of the two presentations (labelled "random re-acquisition + mean of two"), but by using only one presentation. This indicates that, by better choosing the data used for verification, we can reach the same reduction in error rate as would be achieved by using double the amount of data with intra-modal fusion. The advantage of the reliability approach can be pushed further by not throwing away the unreliable presentation completely, a sensible approach since it would have already been re-acquired at this stage.

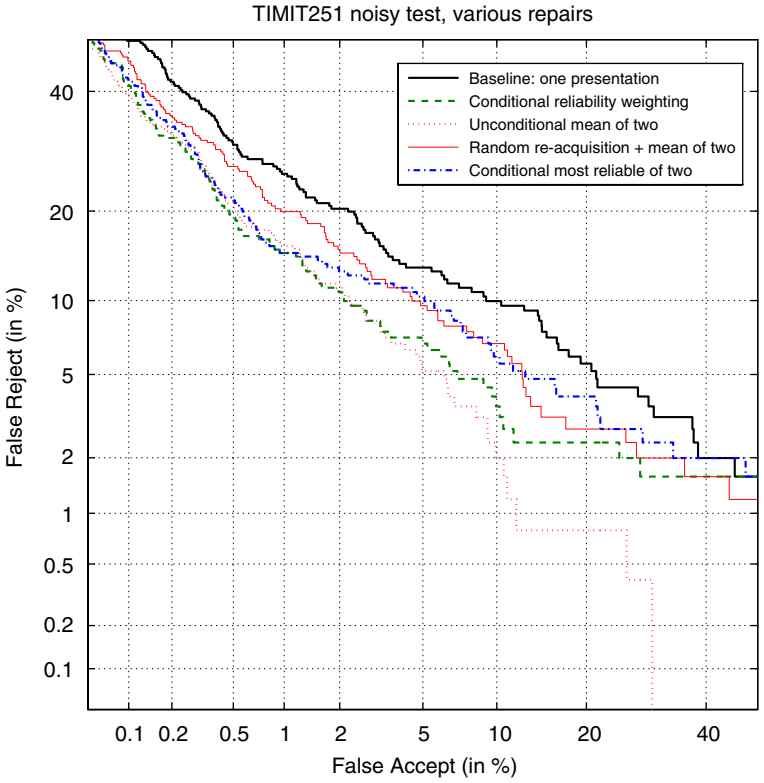

Fig. 7. DET curves for the speaker verification system. The baseline curve is for the system that always accepts the first presentation. *Conditional reliability weighting* is for a reliability-based weighting (see Eq. (9)) that is applied only if the first presentation has low reliability. Mean of two is for a system that always requests two presentations and computes score as the mean of these two presentations.

Thus, the second repair method is based on Eq. (9): if the first presentation is deemed too unreliable, a second one is requested but this time the two scores are fused using a weighted sum where the weighting coefficients are the scaled reliabilities of each presentation. The results for this method are shown in Fig. 7 under the label "conditional reliability weighting". In this case it can be seen that the results are very close to what is reached by a baseline method (labelled "unconditional mean of two") that always requests two presentations. By using this method, the EER is lowered from 10% for the baseline to 6% for the reliability-based repair sequence, a relative reduction of about 40%.

Another possibility would be to perform inference on the *TID* node (see Fig. 3(b)) and compute the posterior $P(TID|Sc, CID, QM)$. This posterior could then be used directly in intra-modal fusion, using for example fixed rules.

Depending on usability factors or security concerns, the reliability threshold can be lowered (so that the system re-acquires more often) or raised to achieve a balance in terms of performance, complexity, and time needed for verification.

6. Conclusions

We have presented confidence measures for speaker verification, and expanded the family of multiple-domain confidence measures by adding a probabilistic measure of decision reliability in speaker verification which has a probabilistic interpretation, takes into account signal-domain auxiliary information, and information about the speaker verification classifier behaviour. Bayesian networks are used to model the dependencies between these sources of information in order to infer the a posteriori distribution over the possible reliability measure values. We showed that the setting of priors is a very important aspect of design in speaker verification, and that Bayesian networks offer a flexible framework to do so.

We compared the performance of confidence and reliability measures on a subset of the TIMIT database comprising 251 users, to which babble-type noise was added in random proportions from 5 to 55 dB. We introduced evaluation criteria and exposed a concern that is particular to the confidence/reliability estimation in biometric identity verification cases: the "double imbalance" of clients versus impostor attempts, and of correct versus incorrect decisions.

The reliability measure was then applied to a speaker verification task to manage a repair sequence with two approaches, one that discards the unreliable presentation and the other which performs intra-modal fusion using reliability as fusion weights. It was shown that error rates can be attained which are close to a system that always requests two presentations, while only re-acquiring data for about 50% of the users.

Future work will be on applying the reliability approach to non-biometric synthetic data and further exploring the multimodal biometric verification case.

Acknowledgements

The authors are grateful to Prof. Athina Petropulu for the invitation to write this article. The authors also acknowledge the insightful remarks of the anonymous reviewer.

References

[1] J.P. Openshaw, J.S. Mason, On the limitations of cepstral features in noise, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 1994, pp. 49–52.
[2] F. de Wet, J. de Veth, L. Boves, B. Cranen, Additive background noise as a source of non-linear mismatch in the cepstral and log-energy domain, Comput. Speech Lang. 19 (1) (2005) 31–54.
[3] Y. Pan, A. Waibel, The effects of room acoustics on MFCC speech parameter, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), 2000.
[4] J. Pitrelli, M. Perrone, Confidence modeling for verification post-processing for handwriting recognition, in: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR), Niagara-on-the-Lake, Canada, August 2002, pp. 30–35.
[5] J. Luo, M. Boutell, Automatic image orientation detection via confidence-based integration of low-level and semantic cues, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 715–726.
[6] Y. Huang, Y. Li, Prediction of protein subcellular locations using fuzzy k-NN method, Bioinformatics 20 (1) (2004) 21–28.
[7] D. Lo, R.A. Goubran, R.M. Dansereau, G. Thompson, D. Schulz, Robust joint audio–video localization in video conferencing using reliability information, IEEE Trans. Instrum. Meas. 53 (4) (2004) 1132–1139.
[8] H. Jiang, Confidence measures for speech recognition: a survey, Speech Commun. 45 (4) (2005) 455–470.
[9] J. Koolwaaij, L. Boves, On decision making in forensic casework, Int. J. Speech Lang. Law: Forensic Linguist. 6 (2) (1999) 164–242.
[10] W.M. Campbell, D.A. Reynolds, J.P. Campbell, K.J. Brady, Estimating and evaluating confidence for forensic speaker recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, 2005, pp. 717–720.
[11] E. Mengusoglu, Confidence measure based model adaptation for speaker verification, in: Proceedings of the Second IASTED International Conference on Communications, Internet, and Information Technology, Scottsdale, USA, November 2003, pp. 286–290.
[12] N. Poh, S. Bengio, Improving fusion with margin-derived confidence in biometric authentication tasks, in: Fifth International Conference Audio- and Video-Based Biometric Person Authentication (AVBPA), 2005.
[13] H. Gish, M. Schmidt, Text-independent speaker identification, IEEE Signal Process. Mag. 11 (4) (1994) 18–32.
[14] H. Nakasone, S.D. Beck, Forensic automatic speaker recognition, in: Proc. ISCA workshop. on speaker recognition—2001: a Speaker Odissey, 2001.
[15] S. Bengio, C. Marcel, S. Marcel, J. Mariethoz, Confidence measures for multimodal identity verification, Inf. Fusion 3 (4) (2002) 267–276.
[16] E. Erzin, Y. Yemez, A.M. Tekalp, Multimodal speaker identification using an adaptive classifier cascade based on modality reliability, IEEE Trans. Multimedia 7 (5) (2005) 840–852.
[17] Q. Li, B.-H. Juang, C.-H. Lee, Automatic verbal information verification for user authentication, IEEE Trans. Speech and Audio Process. 8 (5) (2000) 585–596.
[18] M.C. Huggins, J.J. Grieco, Confidence metrics for speaker identification, in: Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP), 2002.
[19] J. Richiardi, P. Prodanov, A. Drygajlo, A probabilistic measure of modality reliability in speaker verification, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2005, Philadelphia, USA, March 2005, pp. 709–712.
[20] J. Richiardi, P. Prodanov, A. Drygajlo, Speaker verification with confidence and reliability measures, in: Proceedings of the 2006 IEEE International Conference on Speech, Acoustics and Signal Processing, Toulouse, France, May 2006.
[21] K. Murphy, Dynamic Bayesian networks: representation, inference and learning. Ph.D. Thesis, University of California, Berkeley, July 2002.
[22] K. Toyama, E. Horvitz, Bayesian modality fusion: probabilistic integration of multiple vision algorithms for head tracking, in: Proceedings of the Fourth Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, January 2000.
[23] R. Barandela, R.M. Valdovinos, J.S. Sánchez, F.J. Ferri, The imbalanced training sample problem: under or over sampling?, in: Proceedings of the SSPR & SPR 2004, Lecture Notes in Computer Science, vol. 3138, Springer, Berlin, January 2004, pp. 806–814.

[24] D. Reynolds, A Gaussian mixture modeling approach to text-independent speaker identification, Ph.D. Thesis, Georgia Institute of Technology, Atlanta, USA, 1992.
[25] E.K. Patterson, S. Gurbuz, Z. Tufekci, J.N. Gowdy, Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus, EURASIP J. Appl. Signal Process. 2002 (11) (2002) 1189–1201.
[26] D.K. Freeman, G. Cosier, C.B. Southcott, I. Boyd, The voice activity detector for the pan-European digital cellular mobile telephone service, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, 1989, pp. 369–372.
[27] P. Renevey, A. Drygajlo, Entropy based voice activity detection in very noisy conditions, in: Proceedings of the Seventh European Conference on Speech Communication and Technology (EUROSPEECH), 2001.
[28] Z. Chen, X. Ding, Rejection algorithm for mis-segmented characters in multilingual document recognition, in: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2003.
[29] A. Drygajlo, D. Meuwly, A. Alexander, Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition, in: Proceedings of the Eighth European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, September 2003, pp. 689–692.
[30] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, J. Ortega-Garcia, Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition, Comput. Speech Lang. 20 (2–3) (2006) 331–355.
[31] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the International Conference on Machine Learning (ICML), 1997, pp. 179–186.
[32] National Institute of Standards and Technology. The 2001 NIST evaluation plan for recognition of conversational speech over the telephone, October 2000.
[33] J.-F. Bonastre, F. Wils, S. Meignier, ALIZE, a free toolkit for speaker recognition, in: Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, USA, March 2005, pp. 737–740.