A PROBABILISTIC MEASURE OF MODALITY RELIABILITY IN SPEAKER VERIFICATION

Jonas Richiardi[†], *Plamen Prodanov*[‡], *Andrzej Drygajlo*[†]

[†]Signal Processing Institute / [‡]Autonomous Systems Laboratory
Swiss Federal Institute of Technology Lausanne

ABSTRACT

In this paper, a probabilistic measure for reliability of speaker verification under noisy acoustic conditions is proposed. A Bayesian network is used to estimate a probability for verification errors, given the GMM-based speaker verification system output and additional information about the level of acoustic noise. In particular, the log-likelihood ratio and a signal-to-noise related feature are used to account for the adverse acoustic conditions. The probabilistic measure is subsequently employed in governing a repair sequence of trials for acquiring additional speech presentations which are less likely to lead to unreliable verification. The potential of the proposed method is tested through cross-validation experiments. Finally, the benefits of the repair sequence in terms of verification accuracy is evaluated on a noisy environment speaker verification task.

1. INTRODUCTION

The goal of biometric identity verification is to assert whether a certain person is indeed whom she claims to be, based on behavioural or biological traits, also known as modalities. Speech is a personal trait that can be used for biometric user verification and benefits from ease of use for the end user and cheap sensor hardware, and can be deployed in a variety of environments. Speaker verification performance is however very dependent on environmental acoustic conditions, in addition to intra-speaker variability which is strongly influenced by health factors, emotional state, and inter-session time.

It is well known that the statistical distribution of common speech features such as MFCCs is significantly distorted when the original clean speech signal is subjected to additive noise [1]. 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, hence increasing false accept (FA) or false reject (FR) rates.

It has been shown that combining scores for multiple presentations of the same biometric modality (for example face images) reduces the overall verification error rate [2]. Other approaches have combined repeated measurements of the same modality, multiple classifiers on the same modality, and different modalities in a sequential fashion, exploiting the combination of the above strategies until the system performance is sufficient [3].

Accurate verification depends on good data quality. Consequently, the output from the classifier of a speaker verification system may not be sufficiently robust against a noisy environment, and a reliability measure incorporating information about the acoustic environment and the classifier behaviour should be integrated into the overall verification scheme.

This reliability measure can be used in several ways: in the case of a multimodal biometric system it can be used to weigh the speech modality classifier output with respect to other modalities immune to acoustic noise. In a single-modality speech-based biometric system, the reliability measure can be used in a sequential manner to reject presentations for this measure is too low, or to weigh the classifier output scores for several presentations before combining them. It should be noted that in our case the input to the "fusion" decision rule would not directly be the classifier score as is the case in fixed rules [4], but the modality reliability measure obtained through the training process. Recent research in this direction has explored the use of quality measures for improving speaker verification accuracy, using either signal features to weigh a score in verification [5], or score-related quantities [6] for enrollment quality measure.

Our approach combines information about both the acoustic environment and the classifier behaviour in order to provide modality reliability information in 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 modality reliability measure, given additional evidence about the enviromental acoustic conditions. Since the intercausal relations of these two factors cannot be established deterministically, the proposed probabilistic reliability measure is justified. Instead of being assigned a particular value, a probabilistic reliability measure is defined by a distribution over its possible values. In this paper we report on the use of Bayesian networks for inferring the the modality reliability distribution in a speaker verification task under noisy acoustic conditions.

The paper is structured as follows: Section 2 describes Bayesian networks for modelling the distribution of the modality reliability mesure. Section 3 demonstrates the potential of the method through cross-validation experiments. In Section 4 the probabilistic measure is applied to a speaker verification task, in repair sequences for acquiring speech presentations which are less likely to lead to unreliable verification.

2. A BAYESIAN NETWORK FOR SPEECH MODALITY RELIABILITY MEASUREMENT

Bayesian Networks (BNs) are graphical models used to describe a joint probability distribution (pdf) over a finite set of random variables [7], and are completely defined by the triple (V, A, CPD), where V is the set of nodes associated with the random variables,

A is the set of arcs and CPD 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 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.

In our case, in order to represent the real state of the user identity and the verification result we introduce two binary variables: True User identity (TID) and Classified User identity (CID). $TID = 1$ 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 speech-based classifier accepts the identitity claim" ($CID = 1$) and "the speech-based classifier rejects the identity claim" ($CID = 0$). To define the reliability measure we introduce another binary variable MR, where $MR = 1$ represents that the "modality is reliable" and $MR = 0$ represents the opposite statement.

The Bayesian network in Fig. 1 (a) depicts a causal model for the variables TID,CID and MR. In this network the True User Identity can be seen as the cause of a particular Classified User Identity value, and the Modality 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 $MR = 1$ (the classifier makes a correct verification decision because the modality is reliable and the user is a client) or $TID = 0$ and $MR = 0$ (the classifier makes a wrong verification decision even though the user is an impostor because the modality is unreliable).

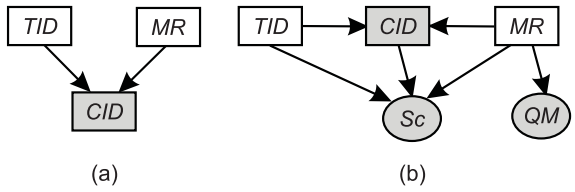

Fig. 1. Bayesian networks for estimation of modality reliability

In this case, $V = (TID, CID, MR)$, and taking into account the arcs defined in Fig. 1(a), the joint pdf over V can be written as:

$$P(TID, CID, MR) = P(MR)P(CID|MR, TID)P(TID)\cdot \quad (1)$$

Since the variables TID and MR 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, MR) values. The verification score (likelihood ratio) is known to carry information about the state of the user identity (client/impostor), while a signal quality measure can be used to provide evidence for the MR variable. For example the signal-to-noise (SNR) ratio 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 Quality Measure for the given modality (SNR in the case of speech). MR, CID and TID can be seen as causes for the observed Sc value, while MR can be seen as the cause for QM values.

The final version of the BN incorporating all these variables $V = (TID, CID, MR, Sc, QM)$ is depicted in Fig. 1 (b). Discrete variables are drawn as squares, and circles are used for the continuous ones. The CPD of a discrete variables is represented by a probability table, while continuous variables make use of arbitrary parametric CPDs. In this paper, we use conditional Gaussian distributions.

In our case the posterior $P(MR|CID, Sc, QM)$ is the distribution of modality reliability measure. To mark an observed variable in the Bayesian network we use shading, and unobserved variables are left white. Once the CPD functions for all the nodes given their parents are defined, an 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 [7].

3. EXPERIMENTS - MODALITY RELIABILITY

3.1. BN Training strategy

In order to perform consistent inference on $P(MR|CID, Sc, QM)$, 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 a maximum likelihood (ML) training technique [7]. After the training, the posterior distribution $P(MR|CID, Sc, QM)$ can be used as a probabilistic mesure of speech modality reliability.

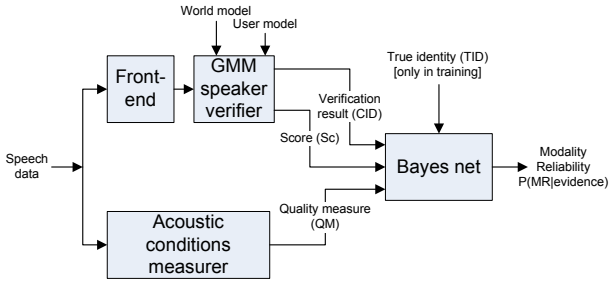

Fig. 2. Combined speaker verification and modality reliability estimation systems

The training setup for the Bayesian network is depicted in Fig. 2. A speaker verification system provides values for the CID, Sc variables, and an acoustic environmental condition measure provides the quality measure (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 about 12 s of held-out data for each user. 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 MR in the BN training data as being "true" ($TID = CID$ then $MR = 1$) or "false" ($TID \neq CID$ then $MR = 0$). In that way, we aim to model the

relationship between verification errors and environmental conditions. Thus, we assume the speaker verification classifier performs above chance level in clean conditions.

3.2. Speaker verification system

The database used for experiments is a 258-users subset of TIMIT, divided in 186 males and 72 females. Approximately 20 seconds of data in 6 presentations (phonetically rich read sentences) for each user is used to train the user models. The remaining 4 client presentations per user are held out for modality reliability experiments. In verification, no separation is made between male and female pool.

The speaker verification system uses voice activity detection to remove pause-related portions of speech, after which 12 MFCCs with first and second order time derivatives are extracted with cepstral mean normalisation. The features are modelled by 64 Gaussian components models with diagonal covariance matrices. Log-likelihood ratio scores are produced using a 64 Gaussians world model for normalisation.

Speaker-dependent thresholds are obtained a priori in the following manner: the 6 presentations used for model training are used to obtain 6 client scores. 100 randomly selected impostors presentations are used to produce 100 impostor scores. Since the client presentations used to obtain verification scores were already used for training, the client scores will be over-optimistic and will result in a very high number of false rejects. This is adressed by assuming that the scores of the 2 best impostor presentations are close to the worse client scores that will be obtained in testing, and adding the former to the client scores pool. Then, a Gaussian distribution is estimated from "impostors only" and "clients+2 best impostors" scores and the intersection point between the curves is taken as threshold. This has the disadvantage of overestimating the variance of the client distribution.

3.3. Acoustic environmental conditions quality measure

To measure the acoustical conditions for the speech modality we use a signal-to-noise ratio (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 in our case we have a single channel speech signal we estimate these energies based on a voice activity detection (VAD) and subsequent pause/speech segmentation. The VAD algorithm is based on the "Murphy algorithm" described in [8]. We then assume that the average energy of pauses is associated with that of noise. Our SNR-related modality quality measure (QM) is given by the formula:

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

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.

3.4. Dataset balancing

Assuming the speaker verification classifier performs well, the data set used for BN training will by definition always contain less data labelled $MR = 0$ ($TID \neq CID$) than data labelled $MR = 1$. 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.

During training of the Bayesian network the prior probabilities over the MR and TID variables are assigned according to the counts of data samples corresponding to the different MR or TID variable realisations. Thus, if the Bayesian network is trained without special care, the learned prior probabilities on the MR and TID variables will be mismatched for client and impostor access because the counts for impostor accesses will be much higher. Since we do not want to bias the final posteriors over the MR 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 MR is a competing cause for the explanation of CID it is also important to balance the number of examples with respect to the MR values.

Therefore the portion of the noisy TIMIT database dedicated to the BN training and testing is balanced accordingly: 876 sequences of the form MR, TID, CID, Sc, QM are used, uniformly distributed with respect to the MR and TID values.

3.5. Experimental results

To test the BN accuracy in predicting modality reliability we perform a series of cross-validation experiment using the balanced database described above. The experiments are done as follows: 1) generate a balanced training data set containing a random selection of 2/3 of all the database examples, 2) use the remaining part as testing data, 3) train and test the BN with the corresponding training and testing data set, 4) iterate the process 100 times. During testing TID and MR variables are unobserved, while CID, Sc, QM are observed (Fig. 1). In these conditions, acccording to the d-separation rules [9] TID and MR become dependent through their common observed children CID and Sc. We calculate the posterior $P(MR|CID, Sc, QM)$ using the junction tree algorithm. To select the most likely value for MR we use an argmax criterion.

This value is then matched over the test set MR labels and accuracies are calculated accordingly. Fig. 3 shows a graphical representation of the posterior $P(MR|CID, Sc, QM)$ resulting from the last experiment iteration. The first graph shows the labeling of the 292 testing examples, where the y-axis value of one corresponds to $MR = 1$, and the y-axis value of zero to $MR = 0$. The second graph corresponds to the hard decision made on MR by the $argmax$ criterion. The third graph depicts the values for $P(MR = 1|CID, Sc, QM)$.

Table 1 presents the results for the client and impostor accuracy along with their standard deviation over 100 runs.

100 trials	clients	impostors	overall
mean accuracy	82.6 %	76.2%	79.4%
σ	2.9%	2.6%	2.1%

Table 1. Accuracy of modality reliability estimation for clients and impostors

4. EXPERIMENTS - IMPROVING SPEAKER VERIFICATION PERFORMANCE

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

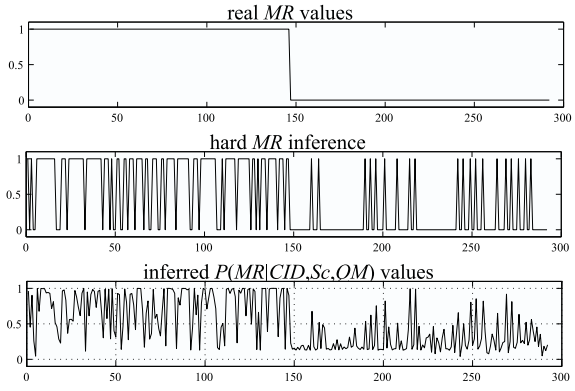

Fig. 3. Graphical representation for $P(MR|CID, Sc, QM)$

speaker verification systems. The baseline system described in subsection 3.2 is forced to take all testing presentations and give a score and accept/reject decision (CID). The improved system implements a simple repair sequence: it is allowed to request another presentation in substitution if it estimates that the modality is unreliable for this presentation ($P(MR = 1|CID, Sc, QM) < 0.5$). If the second presentation also has low reliability, the system picks the presentation with highest modality reliability to produce the final score and accept/reject decision. In this case the behaviour is similar to a max rule on MR for intra-modality fusion.

For the experiment the Bayesian network was trained as described in subsection 3.1, using 248 examples of noisy client presentations and 248 examples of noisy impostor presentations. The test set held out to test the baseline and improved speaker verification systems consists of 258 client presentations and 12900 impostor presentations. In addition, the improved system had another pool of 258 client presentations and 12900 impostor presentations available for re-presentation requests as described above.

As shown in Fig. 4, incorporating modality reliability information lowers the probability of error over most of the operating range and results in the EER dropping from 9.3% to 2.8%.

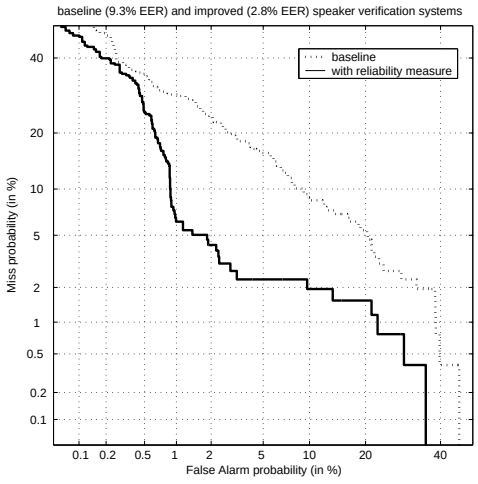

Fig. 4. DET curve for baseline and improved speaker verification systems on noisy data

5. CONCLUSIONS

We have presented a probabilistic measure of modality reliability taking into account a signal-domain measure and information about the speaker verification classifier behaviour. Bayesian networks were used to model the dependencies between these sources of information in order to infer the a posteriori distribution over the possible reliability measure values. It was found that balancing the Bayesian network training set is important in order not to bias the reliability measure in the case of clients or impostors. The reliability measure was then applied to a speaker verification task to manage a repair sequence which requests an additional presentation if the initial presentation is of insufficient reliability. It was shown that a significant gain in terms of verification error rate could be obtained by rejecting low-reliability user presentations according to the reliability measure. Future work includes applying the reliability measure in parallel, rather than sequential, combination strategies, and applying it to the multimodal identity verification problem.

6. ACKNOWLEDGEMENTS

JR wishes to thank Prof. Deutsch of the Acoustics Research Institute in Vienna for his generous short-term provision of working space.

7. REFERENCES

[1] J.P. Openshaw and J.S. Mason, "On the limitations of cepstral features in noise," in *Proc. IEEE ICASSP*, April 1994, pp. 49–52.

[2] J. Kittler, J. Matas, K. Jonsson, and M. Ramos Sánchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters*, vol. 18, pp. 845–852, 1997.

[3] P. Verlinde, P. Druyts, G. Chollet, and M. Acheroy, "Multi-level data fusion approach for gradually upgrading the performance of identity verification systems," in *Proc. SPIE*, 1999, vol. 3719, pp. 14–25.

[4] R. Roli, J. Kittler, G. Fumera, and D. Muntoni, "An experimental comparison of classifier fusion rules for multimodal personal identity verification systems," in *Proc. Third International Workshop on Multiple Classifier Systems*, June 2002, pp. 325–226.

[5] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," in *Proc. Odyssey-04, The ISCA Speaker and Language Recognition Workshop*, 2004, pp. 105–110.

[6] J. Koolwaaij, L. Boves, H. Jongbloed, and E. den Os, "On model quality and evaluation in speaker verification," in *Proc. IEEE ICASSP*, June 2000, pp. 3759–3762.

[7] K. Murphy, "Dynamic bayesian networks: representation, inference and learning," July 2002.

[8] D. Reynolds, *A gaussian mixture modeling approach to text-independent speaker identification*, Ph.D thesis, Georgia Institute of Technology, Atlanta, USA, 1992.

[9] F.V. Jensen, *Introduction to Bayesian networks*, Springer-Verlag New York, 1996.