SPEAKER VERIFICATION WITH CONFIDENCE AND RELIABILITY MEASURES

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

†Signal Processing Institute/‡Autonomous Systems Laboratory
Swiss Federal Institute of Technology Lausanne
{jonas.richiardi,plamen.prodanov,andrzej.drygajlo}@epfl.ch

ABSTRACT

In pattern recognition, the need to quantify the quality of a classifier's output has gained importance in the past years. Speaker verification is no exception. This paper presents a probabilistic reliability framework incorporating signal-domain information into the confidence estimation and contrasts this method with classical approaches to estimating the confidence in a given speaker verification classifier output. We show that the method proposed can deal with adverse acoustic conditions for a wide range of signal-to-noise ratios, does not depend on a Gaussian assumption for impostor and client score distributions, and presents benefits in terms of scalability and interpretability of the measure. We contrast reliability and confidence approaches, and evaluate performance on a degraded version of the 295-users XM2VTS database.

1. INTRODUCTION

Many areas of pattern recognition make use of measures that express the uncertainty in a classifier's output, at the measurement (score) or decision levels. Speech recognition systems often output confidence levels with recognised words, dialogue systems use them to govern actions and repair strategies, and handwriting recognition applications make use of estimates of the segmentation module's output to drive feedback mechanisms. In biometric authentication, in particular speaker verification, the idea that the uncertainty in the output of classifiers should be quantified has gained ground in the last few years.

One of the most important factors affecting the performance of speaker verification systems (apart from lack of training data) is the mismatch between training and testing conditions. Both convolutional (channel) and environmental (partly additive) noise can be present, dramatically impacting recognition performance.

In these circumstances, basing a confidence measure only on classifier-domain information (such as likelihood or distance given by the classifier,posterior probability, n-best lists) is likely to reach limits that require additional information (signal-domain or feature-domain) to overcome. Thus, several approaches have been proposed to try and integrate multiple sources of information into the confidence estimation process ([1], [2], [3], [4]).

In the rest of this paper, we review several approaches for estimating confidence (Section 2), we expand on the theory behind the reliability approach and contrast with confidence measures (Section 3), and present some experiments on the XM2VTS database (Section 4).

2. CONFIDENCE MEASURES IN SPEAKER VERIFICATION

In some speaker recognition applications, such as forensic speaker identification, it is crucial to have an estimation of the certainty of a score. Thus, early research into confidence measures for speaker recognition has originated in forensic science [5].

In this section, we divide confidence measures for speaker verification into two areas: those that use classifier-domain information only (such as scores and posterior probabilities), and those that take into account other source of information such as signal-domain quantities (e.g. signal-to-noise ratio or duration).

2.1. Classifier-domain confidence measures

Knowing the expected impostor and client score distributions provides important insights into the classifier's behaviour. It will be possible to assign confidence values to different portions of the score range. Several methods are used to model confidence in relation with score distributions, of which we provide a brief overview below.

We reformulate Nakasone and Beck's [5] Bayesian confidence measure definition (which matches Fredouille et al's [6] definition for MAP normalisation in a multiple classifier speaker verification system) in our biometric identity verification terms as follows:

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

where TID (true identity) is a binary variable indicating whether the utterance comes from a true client ($TID = 1$) or an impostor ($TID = 0$), and Sc is a continuous random variable indicating the output of the classifier, generally in the form of a log-likelihood ratio. The confidence measure can be phrased as "the *a posteriori* probability that the utterance is from a client given the score". Assuming that the client and impostor score distributions are Gaussian, they then define the confidence measure by fitting a logistic function to the posterior probability represented by Eq. 1:

$$CM_1(Sc) = \frac{e^{(\beta_0+\beta_1 Sc)}}{1+e^{(\beta_0+\beta_1 Sc)}} \quad (2)$$

Bengio et al. [7] proposed three methods for computing confidence, all based on classifier-domain quantities. The simplest method which we present below assumes that impostor and client scores are normally distributed (another non-parametric method is presented in the paper that makes no such assumption), and defines confidence as the difference in probability for a given score between the client and impostor distributions learned on an evaluation set:

$$CM_2(Sc) = |P(Sc|TID=1) - P(Sc|TID=0)| \quad (3)$$

2.2. Multiple-domains confidence measures

More recently, there has been a surge of interest in taking additional sources of information into account.

Campbell et al. [2] use signal-domain data (utterance duration, channel label and signal-to-noise ratio) in addition to utterance score to estimate a confidence for each score. Here, the task is to compare whether two utterances come from the same speaker, using a speaker verification system. The confidence measure in this case can be phrased as "the *a posteriori* probability that the two utterances come from the same speaker". The confidence modelling is done by training a multi-layer perceptron (MLP), meaning no a priori form for the distribution has to be assumed; furthemore the output of the MLP should approximate the Bayesian a posteriori confidence of Eq. 1. However, apart from this single posterior output, which has a clear meaning, the parameters and the architecture of the MLP itself are hard to interpret.

Huggins and Grieco [1] have proposed taking 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 train/test utterance duration and signal-to-noise ratio. Their method is based on computing error rates with respect to 7 discrete SNR levels (pink noise mixed in from 6dB to 24dB) for 13 different utterance durations, thus resulting in 91 regression models indexed by train and test utterance duration. The amount of overlap between models can then be added by performing another level of regression on top of the basic duration/SNR combination 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.

Poh and Bengio [4] use the false accept rate for a certain score (taken as threshold) subtracted with the false reject rate for the same threshold:

$$CM_3(Sc) = |FAR(Sc) - FRR(Sc)| \tag{4}$$

The client and impostor distributions are trained on an evaluation set. This is an interesting approach since it takes into account the distribution of errors with respect to a score. Thus, the closer the score is to the decision threshold, the lower the confidence. They then combine this with a speech quality measure to enhance fusion in multimodal biometrics.

3. RELIABILITY MEASURES FOR SPEAKER VERIFICATION

In all confidence measures presented in Section 2, a quantity that plays a central role is $P(Sc|TID)$, that is the likelihood of a score given a client or impostor score distribution. We argue here that it is more interesting to directly use the *probability of making an erroneous decision* given a certain score and other information, as it allows to quantify the competence of the classifier directly.

We first introduce three new variables. CID is a binary variable which indicates the classifier's decision ($CID = 0$ if the classifier decides for impostor access, $CID = 1$ if the classifier decides for client access). DR (Decision Reliability) is also a binary variable which indicates whether the classifier was wrong ($CID \neq TID$) or correct ($CID = TID$) with respect to the ground truth TID. Finally, $\boldsymbol{QM}$ (Quality Measure) is a vector random variable which contains signal-related quantities pertaining to the amount of noise in the signal. With these variables, our definition of reliability measure, "the posterior probability of taking a correct decision given available evidence", can be written as

$$P(DR = 1|CID, Sc, \boldsymbol{QM}) \tag{5}$$

In this section, we compare our approach (more fully exposed in [3]) with others found in litterature and expose its advantages and disadvantages.

3.1. Speech signal quality measures

Additive noise is known to have a negative impact on speaker verification performance. If the speaker verification system has access to a measure of the amount of noise present with the speech signal, better reliability estimates can be obtained. However, the noise estimation process itself is fallible, especially if it relies on explicit speech/pause segmentation of the source signal with a non-robust voice activity detector (VAD). Thus, it is beneficial to have several estimates of the signal quality and to combine them at a later stage.

The first quality measure used in our experiments, QM_1 is related to the signal-to-noise ratio and uses a VAD algorithm based on the "Murphy algorithm" described in [8]:

$$QM_1 = 10 \log_{10} \frac{\sum_{i=1}^{N} Is(i)s^2(i)}{\sum_{i=1}^{N} In(i)s^2(i)}, \tag{6}$$

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.

The second quality measure, QM_2, is a SNR-related estimate, which is calculated using Equation 6. The difference in this case is that we make use of the short-term spectral entropy for assigning values to $Is(i)$ and $In(i)$. The entropy is a measure defined over a probability distribution function (pdf). It measures the peakiness of the pdf and is closely related to the informativeness of the distribution. The spectral entropy is calculated over the short term spectrum values, where the spectral values are normalized to sum up to 1 thus forming a pdf. The spectral entropy is calculated as follows:

$$H(|\boldsymbol{Y}(w,t)|^2) = -\sum_{w=1}^{\Omega} \frac{|Y(w,t)|^2}{\Sigma_{w=1}^{\Omega}|Y(w,t)|^2} \log\left(\frac{|Y(w,t)|^2}{\Sigma_{w=1}^{\Omega}|Y(w,t)|^2}\right), \tag{7}$$

where $|\mathbf{Y}(w,t)|^2$ is the power spectrum for frame t. $H(|\mathbf{Y}(w,t)|^2)$ is maximized when we have white noise and is minimized when we have a pure tone. The application of entropy relies on the assumption that the presence of pitch in speech segments results in a more organized signal (presenting series of peaks in the spectrum) compared with the case of noise (pauses). Thus, the entropy value is higher for pause than speech regions. The algorithm used in this paper is similar to the one presented in [9].

3.2. Graphical models for reliability modelling

Graphical models offer a very expressive and flexible framework for modelling a variety of phenomena. In our case, we use a Bayesian network to represent the joint conditional distribution of the variables of interest. The topology of the Bayesian network is represented on Figure 1, and the rationale behind it is presented in [3]. Round nodes represent continuous random variables, and square nodes discrete random variables. Shaded nodes (evidential variables)

are observed in testing, others are hidden (no value is provided in testing). Continuous random variables are modelled by Gaussian distributions, and discrete random variables by probability tables estimated from counts in training data.

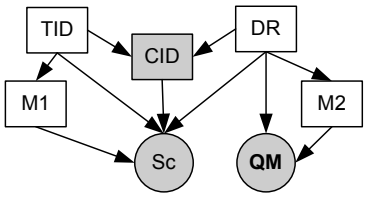

Fig. 1. Graphical model for decision reliability posterior distribution

Given the Bayesian network variables set $V = (DR, TID, CID, Sc, QM, M1, M2)$, where $M1, M2$ represent mixing weights learned through a maximum likelihood algorithm, and taking into account the arcs defined in Fig. 1, the joint pdf over V can be written as:

$$\begin{aligned} P(V) = P(DR)P(TID)P(CID|TID, DR) \\ \cdot P(Sc|DR, TID, CID, M1)P(M1|TID) \\ \cdot P(M2|DR)P(QM|DR, M2) \end{aligned} \quad (8)$$

The posterior $P(DR|CID, Sc, QM)$ is the distribution of the decision reliability measure. Following the network topology the posterior distribution over DR can be written as:

$$P(DR|cid, sc, qm) = \alpha \sum_{TID,M1,M2} P(V) \quad (9)$$

For a given value of DR, say $DR = 1$, and a classifier decision $CID = cid$, the distributive law can be applied to Eq. 9 to simplify the computation:

$$\begin{aligned} P(DR = 1|cid, sc, qm) = \alpha P(DR = 1)P(TID = cid) \\ \cdot \sum_{M1} \underbrace{P(M1|TID = cid)}_{\star} \cdot P(sc|DR = 1, M1) \\ \cdot \sum_{M2} \underbrace{P(M2|DR = 1)}_{\star} P(qm|DR = 1, M2) \end{aligned} \quad (10)$$

The α term is a normalisation coefficient equal to $\frac{1}{P(cid,sc,qm)}$. The term $P(TID = cid)$ is the prior probability on client or impostor access happening. In our case the prior is fixed at 0.5 in training. The $P(DR = 1)$ term is the prior probability of the classifier decision being correct. Since the testing conditions are not entirely known in advance,this is also fixed at 0.5 in training. The terms marked with $\star$ effectively act as mixing coefficients, and the term within the $M2$ summation corresponds to a two-components Gaussian mixture model over the quality measures. The $P(Sc|DR, TID, CID, M1)$ term in Eq. 8 can be simplified to $P(sc|DR = 1, M1)$ in Eq. 10, because a value of 1 for DR means by definition that $TID = CID$, and a certain classifier decision, $CID = cid$ will then be reflected by $TID = cid$. In this case, the term defines a single Gaussian distribution.

Thus, a 2-component Gaussian mixture model is used to model the distribution of scores in the cases of correct reject ($TID = 0, CID = 0$), false accept ($TID = 0, CID = 1$), false reject ($TID = 1, CID = 0$), and correct accept ($TID = 1, CID = 1$). This essentially decomposes the two classical client ($P(Sc|TID = 1)$) and impostor ($P(Sc|TID = 0)$) distributions in four sub-distributions, each having the possibility of deviating from the Gaussian distribution.

The first difference between the reliability approach and classical confidence approaches is that no Gaussian assumption is made about the impostor and client score distributions, since these are modelled by a 2-components GMM. Thus, they are allowed to have different skewness or kurtosis than the normal distribution. However it should be noted that in experiments the difference in prediction accuracy between Gaussian assumption and non-Gaussian assumption for scores proved minute, with very slightly better results for the non-Gaussian assumption. This result is likely to be data and classifier-dependent.

The second difference with classical confidence approaches is that the model is trained explicitly on correct and erroneous classification decisions, meaning that one important quantity is $P(Sc|DR = 1)$, respectively $P(Sc|DR = 0)$, rather than the impostor or client score distributions $P(Sc|TID)$. Thus, the model learns score distributions with respect to classifier behaviour (error or correct classification), and is able to detect these behaviours in testing. The downside is that there are more model parameters to be estimated on evaluation data than for simpler confidence models, resulting in a more time-consuming approach.

An important advantage of the reliability approach is that it readily provides a framework for incorporating several signal quality measures into the reliability estimation problems. It is quite scalable, as adding new quality measures can simply be done by appending to the $\boldsymbol{QM}$ vector. The covariance matrix estimated on that node of the Bayesian network will efficiently model existing correlations between the quality measures, without requiring another layer of modelling for each new information put into the measure.

Lastly, using graphical models for reliability estimation provides a comprehensive interpretation framework, as several posteriors can be elicited from the joint distribution. For example, eliciting $P(DR = 1|CID, Sc, \boldsymbol{QM})$ can be phrased as "the probability of having taken a correct classification decision given evidence", and the posterior $P(TID = 1|CID, Sc, \boldsymbol{QM})$ can be phrased as "the probability of this utterance being a client utterance given evidence", where in both cases "evidence" can be expanded to "the classifier's opinion, the score, and a vector of quality measures".

4. EXPERIMENTS AND RESULTS

The database used in experiments is the XM2VTS database [10], which contains 295 clients, recorded over 4 sessions. The protocol followed is the Lausanne protocol, configuration 1. To investigate the effects of noisy and mismatched conditions, we created a second version of this database by mixing in non-stationary additive babble-type noise recorded in a lively cafeteria in amounts from 20 to 0 dB SNR, where the noise sample is longer than the utterances and a random portion of the noise is chosen each time.

Each of the 200 clients provides 3 evaluation utterances, and 25 evaluation impostor each provide 8 utterances. This results in 600 evaluation client accesses and 40000 evaluation impostor accesses. The total amount of client test accesses is 400, with 112000 impostor accesses. The test results for the speaker verification system on clean and noisy test data are shown in Fig. 2 to provide an overview of the extent of the degradation applied.

The speaker verification system used is based on the Alize toolkit [11]. The Alize speech/pause detector is run to remove silence portions of the input speech signal before feature extraction. Features used are 12 MFCCs with delta and acceleration coefficients,

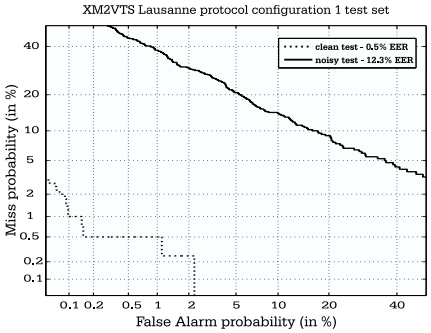

Fig. 2. DET curves for speaker verification system with clean and noisy test data, and in both cases clean training data

and cepstral mean normalisation. A world model is trained from the pooled clean training data of all 200 clients (each providing 3 recordings of about 10 seconds), using 200 diagonal covariance-matrix Gaussian components. Each client's model is then adapted (means only) with their own 3 recordings using MAP adaptation.

The scores resulting from running the system on clean evaluation data are then used to estimate a global threshold T, which is set to the threshold which gives $FAR = FRR$ on the evaluation set (EER threshold).

The testing is then run again on the noisy version of the XM2VTS database. The noisy evaluation data is used to train the parameters of the Bayesian network presented in Section 3.2.

4.1. Accuracy of confidence and reliability measures

To evaluate the benefits brought by the reliability approach with respect to classifier-domain confidence measures, we compute confidence and reliability measures for all tests in the noisy database. Then, we threshold the measures at 0.5: a confidence or reliability above this threshold means the classifier should be trusted, whereas its decision is probably wrong if the measures are below 0.5. We then compare this thresholded confidence or reliability with ground truth labels (TID) and compute accuracies. The results are presented in Table 1. The first thing to note is that, as can be expected, the logistic confidence measure (CM_1) can be improved if the estimates for clients and impostors distributions are trained on a noisy evaluation set. The second thing to note is that factoring noise into the reliability process improves the results significantly. Thus, training confidence measures on noisy evaluation data is a way to improve the accuracy of confidence estimation, but it is better to take into account the amount of noise in the signal explicitly, as is done in the reliability estimation. Thirdly, as expected, the combination of several signal quality measures improves the accuracy of the reliabilty estimates. Lastly, a note of caution is in order, as the accuracy results provided effectively bundle together performance of the measures on false accepts, false rejects, correct accepts, and correct rejects. Depending on prior values and training/testing set structure, the confidence and reliability measures will perform differently in some regions.

5. CONCLUSIONS

We have compared various approaches to automatically evaluate the uncertainty in a classifier's decision, and have shown that in the case of speech it is important to take the environmental conditions into account when mismatch is expected between training and testing environments. We proceeded to highlight the differences between the classical confidence and reliability approaches: easy integration of several signal quality measures, no Gaussian assumption on score or quality measure distributions, training focussed on classifier behaviour, and interpretability of the measure. The reliability measure performed well on a 295-users database to which babble-type noise was added. For biometric authentication tasks, the accuracy is sufficiently high that a decision most likely to be unreliable can be deferred to a human operator when available, a reacquisition can be performed, or a second modality be used. In forensic cases the reliability approach can be used to provide additional information about the performance of the automatic system, if the court can set the prior on TID (which the expert has no right to decide upon).

Further work will include implementation of and comparison with multiple-domain confidence measures.

method	clients	imps.	mean
CM_1 trained on clean data	68.9%	49.7%	59.3%
CM_1 trained on noisy data	85.8%	47.0%	66.4%
CM_3 trained on clean data	50.1%	70.4%	60.3%
CM_3 trained on noisy data	87.1%	22.7%	54.9%
reliability $\boldsymbol{QM} = (QM_1)$	69.8%	92.3%	81.0%
reliability $\boldsymbol{QM} = (QM_1, QM_2)$	73.0%	92.2%	82.6%

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

6. ACKNOWLEDGEMENTS

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in [10]. We also wish to thank Dr. Anil Alexander for his insightful comments.

7. REFERENCES

[1] Mark C. Huggins and John J. Grieco, "Confidence metrics for speaker identification," in *Proc. 7th ICSLP*, 2002.

[2] W.M. Campbell, D.A. Reynolds, J.P. Campbell, and K.J. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proc. ICASSP 2005*, 2005, vol. 1, pp. 717–720.

[3] J. Richiardi, P. Prodanov, and A. Drygajlo, "A probabilistic measure of modality reliability in speaker verification," in *Proc. ICASSP 2005*, Philadelphia, USA, March 2005, pp. 709–712.

[4] N. Poh and S. Bengio, "Improving fusion with margin-derived confidence in biometric authentication tasks," in *Proc. 5th AVBPA*, 2005.

[5] H. Nakasone and Steven D. Beck, "Forensic automatic speaker recognition," in *Proc. 2001: A Speaker Odyssey*, 2001.

[6] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: A block-segmental multirecognizer architecture for automatic speaker recognition," *Digital Signal Processing*, vol. 10, no. 1, pp. 172–197, 2000.

[7] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz, "Confidence measures for multimodal identity verification," *Information Fusion*, vol. 3, no. 4, pp. 267–276, Dec. 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] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proc. EUROSPEECH 2001*, 2001.

[10] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd AVBPA*, 1999, pp. 72–77.

[11] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. ICASSP 2005*, Philadelphia, USA, March 2005, pp. 737–740.