ESTIMATING AND EVALUATING CONFIDENCE FOR FORENSIC SPEAKER RECOGNITION*

W. M. Campbell, D. A. Reynolds, J. P. Campbell, and K. J. Brady

MIT Lincoln Laboratory
Lexington, MA 02420
E-mail: {wcampbell,dar,jpc,kbrady}@ll.mit.edu

ABSTRACT

Estimating and evaluating confidence has become a key aspect of the speaker recognition problem because of the increased use of this technology in forensic applications. We discuss evaluation measures for speaker recognition and some of their properties. We then propose a framework for confidence estimation based upon scores and meta-information, such as utterance duration, channel type, and SNR. The framework uses regression techniques with multilayer perceptrons to estimate confidence with a data-driven methodology. As an application, we show the use of the framework in a speaker comparison task drawn from the NIST 2000 evaluation. A relative comparison of different types of meta-information is given. We demonstrate that the new framework can give substantial improvements over standard distribution methods of estimating confidence.

1. INTRODUCTION

Bayesian methods have become a popular method of approaching speaker recognition in a forensic setting [1, 2]. Some of this success is due to the fact that Bayesian methods aim to produce human interpretable scores. For speaker verification in the Bayesian approach, the starting point is hypotheses, $\omega = 1$ and $\omega = 0$, corresponding to the target speaker present or not present, respectively, and evidence E which has bearing on the hypotheses. The *a posteriori* probability $p(\omega = 1|E)$ is calculated using Bayes rule,

$$p(\omega = 1|E) = \frac{\pi_1 p(E|\omega = 1)}{\pi_0 p(E|\omega = 0) + \pi_1 p(E|\omega = 1)}, \quad (1)$$

where we have used the convention $\pi_i = p(\omega = i)$. The π_i are usually referred to as the priors or the prior probabilities. We call the probability $p(\omega = 1|E)$ the *confidence* in the hypothesis $\omega = 1$.

*This work was sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

The Bayesian approach to inference includes several facets that are interesting and relevant to speaker recognition. First, the Bayesian approach provides a systematic way of incorporating prior knowledge and cost into the decision process. Second, the Bayesian methodology provides an interpretation of probability as degree of belief. This is especially important in selecting priors where one can use subjective or objective approaches. For the latter method, one tries to pick uninformative priors that do not bias a decision toward personal belief [3]. Finally, Bayesians have done an extensive amount of work on evaluating confidence. This work has come about since the Bayesian approach encourages elicitation of probabilities, rather than hard decisions. An introduction to some of the evaluation frameworks is given in [4].

We note that the Bayesian approach is an *interpretation* of probability. Probability theory is rigorously defined axiomatically through Kolmogorov's measure theory methods [5]. For convenience, we use Bayesian language for discussion, but it should be noted that other methods of interpretation (e.g., frequentist) may be more appropriate in some applications.

Our goal in this paper is to consider methods for estimating and evaluating confidence, $p(\omega|E)$. In many cases, E includes a plethora of information available in the recognition process. For the purposes of this paper, we limit ourselves to information automatically determined in the process of speaker recognition. E typically includes scores from several classification systems, such as a Gaussian mixture model (GMM) or Support Vector Machines (SVMs) as well as meta-information typically not explicitly used in the final decision process—channel labels, duration, SNR, etc. Prior work has considered this meta-information for improvement of speaker recognition accuracy, threshold stabilization, and other types of confidence [6, 7, 8]. We show that meta-information can be used in a unified framework for simultaneously improving accuracy and *a posteriori* probability confidence estimation.

The outline of our paper is as follows. In section 2, we review a standard baseline approach to estimating con-

fidence. Next, section 3 discusses methods and provides insight into confidence evaluation. Section 4 provides our framework for estimating confidence. Finally, Section 5 shows experiments using this method on a NIST evaluation.

2. CONFIDENCE ESTIMATION BASELINE

For this paper, we consider a slight variant of the speaker verification task common in forensic work—speaker comparison. That is, given two utterances, we want to find the confidence that the speakers are the same in both utterances—hypothesis $\omega = 1$. From this, we also obtain the probability of the hypothesis, $\omega = 0$, that the two speakers are different.

One approach to this problem is to apply a speaker verification system to the task. For utterance i, we train a speaker model, m_i. We then apply the speaker model m_i to utterance j, $j \neq i$, to obtain a score, s_i. A typical raw score for this system would be the symmetrized score, $s = 0.5s_1 + 0.5s_2$. In order to estimate confidence, a baseline strategy would be to model the score distribution conditioned on a hypothesis as Gaussian; i.e., we have

$$p(s|\omega = i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(s-m_i)^2}{2\sigma_i^2}}. \tag{2}$$

The likelihood ratio, LR, is given by

$$LR = \frac{p(s|\omega = 1)}{p(s|\omega = 0)}. \tag{3}$$

The *a posteriori* probability is then a slight rearrangement of (1),

$$p(\omega = 1|s) = \frac{1}{1 + (\pi_0/\pi_1)(1/LR)}. \tag{4}$$

There is some ambiguity in this approach, since we have to determine which score to model as Gaussian. A natural candidate for a GMM is the average log likelihood ratio of the target speaker score to the universal background model score [9]. This score has a range of $(-\infty, \infty)$ and is produced by a sum of many (assumed) independent random variables—the log-likelihood ratio score at each frame.

This simple approach can be improved by more advanced distributional modeling—e.g., Parzen estimators or GMMs of the score distribution. Since the score distributions are known to be non-Gaussian (e.g., [10]), this will certainly improve estimation of LR. A difficulty arises when we want to incorporate meta-information into the confidence estimation process—what distribution do we select to model the joint distributions between the score and meta-information? Rather than solve this problem, we propose a discriminative approach in section 4.

3. EVALUATING CONFIDENCE

The overall goal of confidence evaluation is to rank different confidence estimators using a numerical measure of goodness. The problem has been studied in statistics [11, 12], meteorology [13], and in speech processing [4, 14].

Confidence evaluation is closely related to confidence elicitation [11] and Bayesian methods. Suppose a weather forecaster gives a certain probability of rain tomorrow. How can we be sure that the forecaster is giving his best estimate of probability according to his beliefs and not hedging to improve some other criterion (e.g., salary)? Another question is, if there are multiple forecasters, then who is the best?

A solution to the problem presented is to use *strictly proper scoring rules*. A scoring rule is a method of assessing the quality of a forecaster. That is, suppose that given the evidence, the forecaster gives a confidence estimate $q(E)$, then a scoring rule assigns a number $M(q)$ which reflects the quality of the estimate. Since E and ω are random variables, $M(q)$ nominally depends on the actual distribution $p(E, \omega)$. A strictly proper scoring rule is one for which the *only* maximizing value of $M(q)$ is when $q(E) = p(\omega = 1|E)$. Thus, a strictly proper scoring rule encourages the forecaster to elicit his personal beliefs.

Numerous strictly proper scoring rules exist. We choose a rule based upon information theoretic measures—the normalized cross entropy metric (NCE). NCE measures the relative reduction in uncertainty over a baseline. For the case when the baseline is the entropy of the hypothesis, we have

$$\mathrm{NCE}(q) = \frac{H(\omega) - \mathrm{E}[-\log_2 |q(E) + \omega - 1|]}{H(\omega)}, \tag{5}$$

where $H(\omega)$ is the entropy of ω and $\mathrm{E}[\cdot]$ denotes expectation with respect to $p(E, \omega)$. $H(\omega)$ represents the baseline where the only information we have is the match prior, π_1.

Calculating NCE in (5) is straightforward. We take the true trials and false trials and calculate

$$\begin{aligned} H_{\mathrm{cond}} = &- \pi_1 \frac{1}{N_{\mathrm{t}}} \sum_{i=1}^{N_{\mathrm{t}}} \log_2(q(E_i^{\mathrm{t}})) \\ &- \pi_0 \frac{1}{N_{\mathrm{f}}} \sum_{i=1}^{N_{\mathrm{f}}} \log_2(1 - q(E_i^{\mathrm{f}})), \end{aligned} \tag{6}$$

where the first sum in (6) is over true trials and the second sum is over false trials. Then NCE is

$$\mathrm{NCE}(q) = \frac{h(\pi_1) - H_{\mathrm{cond}}}{h(\pi_1)}, \tag{7}$$

where $h(\cdot)$ is the entropy function $h(p) = -p\log_2(p) - (1-p)\log_2(1-p)$.

The optimal (maximum) value for NCE occurs when $q(E) = p(w = 1|E)$. Then,

$$\mathrm{NCE}(p) = \frac{H(\omega) - H(\omega|E)}{H(\omega)}, \tag{8}$$

so the maximum value of NCE corresponds to the relative reduction in uncertainty in bits when E is known.

It is important to note that NCE is sensitive to the scale of the confidence. If the confidence is close to 0 for a true trial or close to 1 for a false trial, a large penalty is incurred. Also, if a confidence estimator gives values close to 0.5 all of the time, then $H_{\text{cond}} = 1$, which makes NCE ≤ 0.

Another interesting baseline is to evaluate the NCE for a hard decision approach. Suppose we know the threshold for the equal error rate (EER). Then given a score above the threshold, we produce a confidence of (1-EER). Below the threshold, we produce a confidence of EER. Then, a simple calculation shows that H_{cond} in (6) is $h(\text{EER})$. We could use this in (5) instead of $H(\omega)$, if we wanted a more informative baseline. A key point of exploring the hard decision baseline is that it clearly shows that NCE is influenced by the accuracy of the system. Thus, improvements in NCE can be made by more accurate modeling of the distributions in the LR (3) or by just improving system accuracy.

4. A FRAMEWORK FOR CONFIDENCE ESTIMATION

For confidence estimation in the speaker comparison problem described in Section 2, there are many sources of information. Scores from the train/test process are available, s_1 and s_2. Scores from other classifiers may be available. In addition, various meta-information is available—duration of the utterances, channel labels (either automatically or manually generated), SNR estimates, and numerators and denominators of log likelihood ratios. All of this information contributes to confidence estimation.

To perform confidence estimation using all of the mentioned sources of information, we propose using a multilayer perceptron (MLP), see Figure 1. Inputs to the MLP, $\mathbf{x}$, include the scores and meta-information previously described. The model parameters of the MLP (including the biases) are specified by a vector, $\mathbf{w}$.

The MLP is optimized using a training set labeled with truth. We train the MLP using the same cross-entropy criterion as the NCE. A well-known result [15] is that training with this criterion yields approximations to the *a posteriori* probability, $p(\omega = 1|\mathbf{x})$. Thus, the output of the MLP, $y = f(\mathbf{x}, \mathbf{w})$, estimates the desired probability, $p(\omega = 1|E)$, assuming that the evidence is $\mathbf{x}$.

As an aside, we mention that MLP training criteria and strictly proper scoring rules are closely related. Applying a cross-entropy training criterion to the MLP training "encourages" it to elicit confidences.

We use the Netlab tool [16] for training the MLP; a scaled conjugate gradient algorithm is used for optimization. We ensure that the training priors are controlled by sampling with replacement from the training set. We convert the output of the MLP to LR using the equation

$$LR = \frac{\pi_0^{\text{train}}}{\pi_1^{\text{train}}} \frac{p(\omega = 1|\mathbf{x})}{1 - p(\omega = 1|\mathbf{x})}, \tag{9}$$

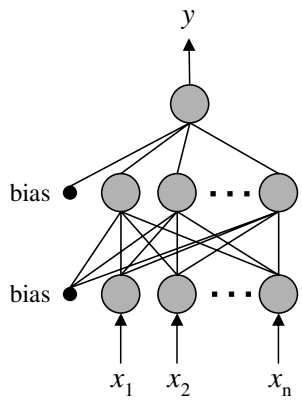

Fig. 1. Multi-layer perceptron for confidence estimation

where the superscript *train* indicates the training priors. The likelihood ratio, LR, in (9) can be converted back to a confidence estimate using the desired priors, π_i, and (4). A convenient fact of this approach is that if the testing priors change, we do not have to retrain the MLP. As a final processing step, we limit the posterior probability to $[0.01, 0.99]$.

Symmetry may be a concern when using the MLP. That is, if the role of the compared utterances is swapped, then the confidence should be the same. One convenient way of dealing with this issue is to evaluate the MLP twice—once with the inputs in one order and then swapped—and average the two outputs. We found this typically improved the quality of confidence estimation. Alternately, symmetry can be incorporated in the training process.

5. EXPERIMENTS

For evaluation, we used the male subset of $3,483$ files from the training and testing portion of the NIST 2000 speaker recognition evaluation (which uses Switchboard 2 phases 1 and 2). This evaluation setup gave a set of durations nominally around 30 seconds and 2 minutes. Since our task was to perform speaker comparison, we scored all possible combinations of files, resulting in $17,776$ true trials and $6,046,127$ false trials.

A GMM verification system was used to produce scores. Front-end processing included 19 MFCCs plus deltas, RASTA, feature mapping [17], and mean and variance normalization. A 2048 component mixture model was used. Training was accomplished by Bayesian adaptation of the means [9]. No additional score normalization was performed by the system (such as Tnorm).

Training for the MLP system was drawn from the Switchboard 2 phase 3 corpus; note that this is a different subset of Switchboard than the NIST 2000 evaluation. Utterances were randomly truncated to give diverse durations. Channel labels were determined automatically in the feature mapping process [9] and consisted of carbon button, elec-

Table 1. MLP inputs and dimension for various configurations

MLP1	$s_{\text{avg}} = 0.5s_1 + 0.5s_2$	1
MLP2	s_i	2
MLP3	s_i, duration	4
MLP4	s_i, duration, num and den of GMM LR	8
MLP5	s_i, duration, channel type	10
MLP6	s_i, duration, channel type, SNR	12

Table 2. Comparison of NCE in percent for different types of meta-information. Results are relative change in uncertainty with respect to a Gaussian distribution baseline

π_1	MLP1	MLP2	MLP3	MLP4	MLP5	MLP6
0.01	9.9	10.6	13.2	11.1	14.3	11.9
0.10	15.0	15.6	17.4	14.3	19.6	16.2
0.25	16.1	16.7	18.4	14.7	20.8	17.4
0.50	16.4	17.0	19.0	14.8	21.4	18.3
0.75	16.3	17.0	19.6	14.9	21.5	19.3

tret, and digital cell. SNR was determined using the NIST *stnr* tool.

Table 1 shows the various configurations of data supplied to the MLP. The MLP was trained with 10 hidden units except for MLP6, which had 12 hidden units (all have one hidden layer). The number of hidden units was determined using a held-out portion of the training set. Note that channel information was discrete and coded in binary form as $\text{cb} = (0, 0, 1)$, $\text{elec} = (0, 1, 0)$, $\text{digital} = (1, 0, 0)$.

Results for the various configurations are shown in Table 2. As a baseline for the NCE in (5), we used a Gaussian-distribution-based approach as described in Section 2 rather than $H(\omega)$. We mention that at $\pi_1 = 0.5$, the distribution-based method gives a $H_{\text{cond}} = 0.489$.

Table 2 shows that a non-Gaussian assumption (MLP1) and duration (MLP3) improve performance substantially over the base. Slightly helpful are the addition of individual scores (MLP2) and channel labels (MLP5) to the MLP's inputs. Finally, both splitting the GMM likelihood ratio into numerator and denominator (MLP4) and using SNR (MLP6) seems to degrade confidence estimation. Overall, we see the inclusion of meta-information substantially improves confidence estimation.

As mentioned in Section 3, accuracy has an impact on the NCE. If we evaluate the distribution method with a hard decision confidence, then $h(EER) = h(0.1193) = 0.527$. For the best performing system, MLP5, at $\pi_1 = 0.5$, $H_{\text{cond}} = 0.384$ and $h(EER) = h(0.1073) = 0.492$. Thus, if we normalize out the accuracy (i.e., use the hard decision as a baseline), then the distribution system has an NCE relative change of 7%, and MLP5 has a relative change of 22%. In a loose sense, this result shows that the MLP5 is not just improving accuracy of the system, it is also improving the *presentation* by producing better confidences.

6. CONCLUSIONS

We explained and showed examples of evaluating confidence for speaker recognition based upon Bayesian and NCE methods. We demonstrated that an MLP framework including meta-information effectively handled multi-condition confidence estimation with performance improvements of $> 20\%$ in NCE.

7. REFERENCES

[1] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM)," in *Proc. Odyssey01*, 2001, pp. 145–150.

[2] J. Gonzalez-Rodriguez, D. Garcia-Romero, M. Garcia-Gomaran, D. Ramos-Castro, and J. Ortega-Garcia, "Robust likelihood ratio estimation in Bayesian forensic speaker recognition," in *Proc. Eurospeech*, 2003, pp. 693–696.

[3] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, NY, 1985.

[4] N. Brümmer, "Application-independent evaluation of speaker detection," in *Proc. Odyssey04*, 2004, pp. 33–40.

[5] P. Billingsley, *Probability and Measure*, Wiley, 1995.

[6] M. C. Huggins and J. J. Grieco, "Confidence metrics for speaker identification," in *Proc. ICSLP*, 2002, pp. 1381–1384.

[7] N. Mirghafori and M. Hébert, "Parameterization of the score threshold for a text-dependent speaker verification system," in *Proc. ICASSP*, 2004, pp. 361–363.

[8] J. Pelecanos, U. Chaudhari, and G. Ramaswamy, "Compensation of utterance length for speaker verification," in *Proc. Odyssey04*, 2004, pp. 161–164.

[9] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[10] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. ICASSP*, 1997, pp. 1071–1074.

[11] Leonard J. Savage, "Elicitation of personal probabilities and expectations," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 783–801, 1971.

[12] M. H. DeGroot, "The comparison and evaluation of forecasters," *The Statistician*, vol. 32, no. 1/2, pp. 12–22, 1983.

[13] R. L. Winkler and A. H. Murphy, ""Good" probability assessors," *Journal of Applied Meteorology*, pp. 751–758, Oct. 1968.

[14] M. Siu, H. Gish, and F. Richardson, "Improved estimation, evaluation and applications of confidence measures for speech recognition," in *Proc. Eurospeech*, 1997, pp. 831–834.

[15] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[16] I. T. Nabney, *NETLAB Algorithms for Pattern Recognition*, Springer-Verlag, 2002.

[17] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, 2003, pp. II–53–56.