On the Performance and Use of Speaker Recognition Systems for Surveillance

Peter J. Barger, Sridha Sridharan
Speech and Audio Research Laboratory
Queensland University of Technology
Brisbane, Queensland, Australia
p.barger@student.qut.edu.au, s.sridharan@qut.edu.au

Abstract

We model the performance of a speaker recognition system used for surveillance to prioritize a large number of candidate speakers in search of a single target speaker. It is assumed that the system operates by ordering all speakers in order from best match to worst match, with the goal of having the true speaker sample positioned as high as possible on the list. Some performance measures for prioritization systems are given and are applied to a real speaker recognition system. An analytic expression for the probability density function of the true speaker's position on the list is found, subject to basic assumptions concerning the distribution of true speaker and false speaker scores. A comparison is made to the performance of a system which is operating only by making verification decisions, and it is shown that making soft decisions results in significantly better surveillance performance.

1. Introduction

There are two main applications for speaker recognition technology. The most frequently considered is the verification of cooperative speakers on a per-trial basis. The other is for surveillance of a large set of speech samples to try to locate a given speaker of interest. The obvious application is to law enforcement and intelligence activities but areas such as news indexing are also relevant [1]. The analysis presented here is applicable to other similar areas of probabilistic pattern recognition.

Some researchers have concentrated on the open-set identification task, where the question asked of the system is, "which, if any, of the sample speakers is the target speaker?". While this is indeed the question in general to be addressed by surveillance systems, the performance metric universally used is the raw recognition rate. We believe that for some applications a surveillance system that consistently *almost* recognizes the target speaker can be just as useful as one which is consistently correct and present here performance measures to demonstrate this. We were unable to locate previous work discussing this aspect of the use of speaker recognition systems for surveillance.

2. The need for caution

It is widely recognized that caution is required when advocating the use of speaker recognition systems in forensic contexts [4]. This is because: the ramifications of the experiment outcome are potentially severe; it is easy for non-experts to place too much reliance on an experiment outcome without understanding the true significance of the result; the speaker recognition experiment is merely part of a larger decision-making process and the technology must be set up in such a way as to not usurp parts of that process which do not properly belong to it; and the effects of such factors as intentional voice disguise, time lapse, illness and uncontrolled recording conditions are poorly understood at best.

These words of warning apply also to the use of automatic recognition technologies in surveillance systems. In addition, the probable secrecy (with respect to the subjects) of the surveillance operation denies them the opportunity to rebut or refute any claim of identity by the system. This applies to an even greater extent to systems used in the manner outlined in this paper, where substantial numbers of errors are to be expected relative to the number of successful detections, thus necessitating the intervention and cooperation of human experts. It is the responsibility of system developers to educate users and potential users of surveillance systems regarding the dangers of naive use of such systems and to emphasize that surveillance systems should not be used as primary evidence of identity in subsequent actions, legal or otherwise. Nevertheless, surveillance systems are an invaluable tool for directing the attention of suitably trained experts towards subjects who are more likely to be of attention to them, and it is in this context that this paper is written.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

3 Verification mode

Speaker recognition researchers and users are familiar with the verification mode of system use. This is based on a series of individual trials, each of which results in a decision to accept or reject the claimed identity. Using this mode of operation for surveillance means treating each sample as a claim to identity of the target speaker and deciding whether or not to accept the claim based on whether or not the score produced by the system exceeds some pre-set threshold. This threshold may be dependent on the target speaker, and may even vary randomly from trial to trial [3], but the important fact is that a hard accept/reject decision is made.

3.1 Verification behaviour

Verification behaviour is well-captured by the standard Detection Error Tradeoff (DET) curve [5] which shows the relationship between the probability of a miss $p(m)$ and the probability of a false alarm $p(f)$. The DET curve is in essence a ROC curve displayed on axes scaled by normal deviates rather than linearly. An intuitively comfortable statistic of interest for a surveillance system is false alarms per hit, and the associated rate of misses for a given operating point. If N_f false trials must be attempted for every true trial, we find that the number of false alarms is $N_f p(f)$ while the number of successful hits is $1 - p(m)$. This gives the system a performance of $\frac{N_f p(f)}{1-p(m)}$ false alarms per hit, while missing $p(m)$ of the targets.

3.2 Verification example

In [9] there appears an argument that large scale surveillance systems are impractical, as the number of false alarms will inevitably swamp the number of hits, even assuming unreasonably small error probabilities. The example given is as follows. A surveillance system is required to search 10^{12} samples looking for 10 target samples. For simplicity, we reduce this to a system searching 10^{11} samples looking for 1 target sample, though the two tasks are not equivalent, as we will discuss later. The system has a false negative rate $p(m) = 0.001$ and a false positive rate of $p(f) = 0.01$. The author of [9] does the arithmetic described above and concludes, 'This unrealistically-accurate system will generate one billion false alarms for every real terrorist plot it uncovers.'

We argue that this reasoning is valid if the system is operating in verification mode, where a hard yes/no decision is made for each sample. However, in some surveillance applications, the performance can be improved significantly.

4 Prioritization mode

Surveillance systems need not make hard decisions, and in view of the points of expressed in Section 2, it may be unwise to do so. A more appropriate use may be as a tool to prioritize samples for further examination by experts.

It is assumed that the surveillance system is operating on a large set of samples, with the task of interest being to locate the single speaker of interest. While this may seem an unrealistic constraint, it models somewhat accurately a type of surveillance application where the primary goal is to search for a single target, regardless of the number of actual targets present. This may be, for example, because once a single target is located, associated collateral information allows the easy (or easier) detection of all remaining targets.

The system presents its finding in the form of a *queue* - an ordered list of samples in descending order of likelihood of being the speaker of interest. We further assume that the user of the system has no further information on the likelihood of the speaker of interest being in any given sample, but is able, on inspecting the samples, to make a determination as to whether the true speaker is present or not. Thus, the user's rational approach is to inspect each sample in order from the top of the queue down. The surveillance system will be most useful when the target appears as close as possible to the top of the queue. It is clear that assessing prioritization performance by examining verification performance is equivalent to assuming that all target samples scores greater than the threshold are strictly less than all false samples that also exceed the threshold, which is clearly not the case.

However, in reality there may be multiple target samples present, and the system as described will detect the target who is positioned *highest* in the queue. This will tend to improve overall performance under the constraints assumed, so we have concentrated on analysing the average performance as a worst-case scenario.

5 Modelling queueing performance

We assume the the recognition system is some type of system which produces a real-valued score, indicating the likelihood that the sample under examination is a match to the target. Current state-of-the-art speaker recognition technologies use Gaussian mixture models (GMMs) which exhibit this type of behaviour. It is assumed that true speaker scores are distributed as univariate Gaussian variables, as are false speaker scores. False speaker scores are assumed to have been normalized to a standard normal distribution, that is with mean $\mu_f = 0$ and standard deviation $\sigma_f = 1$. True speaker scores are also normally distributed, with mean μ_t and standard deviation σ_t. While it is known that neither true speaker scores or false speaker score are

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

actually normally distributed [6], most researchers assume that they are. This not only provides a benefit in simplicity, it has also been used to develop several effective score normalisation techniques such those in [8, 2].

We assume that there is a set of N_f false speaker samples and one true speaker sample, making the total sample size $N = N_f + 1$.

We are interested in the proportion of false speakers who are expected to score higher than the true speaker. If the false speaker's score was t, then the proportion of false speakers who scored higher than t is equal to the probability that a single speaker scores higher than t. We assume that $\mu_f = 0$ and $\sigma_f = 1$, which gives

$$\rho = P(y > t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left[-\frac{s^2}{2}\right] ds. \quad (1)$$

$$\begin{aligned} \rho &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-s^2}{2}\right) ds \\ &\quad - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(\frac{-s^2}{2}\right) ds \\ &\quad - \frac{1}{\sqrt{2\pi}} \int_{0}^{t} \exp\left(\frac{-s^2}{2}\right) ds \\ &= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{t}{\sqrt{2}}\right) \end{aligned} \quad (2)$$

where

$$\text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds. \quad (3)$$

Since $\rho = u(t)$ is a one-to-one function, there exists an inverse function $t = w(\rho)$, and it is easy to find that

$$w(\rho) = \sqrt{2}\,\text{erf}^{-1}(1 - 2\rho) \quad (4)$$

where $y = \text{erf}^{-1}(x)$ is the inverse function of $y = \text{erf}(x)$.

If the distribution function of t is $f(t)$, and $\rho = u(t)$, ρ has the distribution function

$$g(\rho) = f\left[w(\rho)\right] |J|, \quad (5)$$

where $J = \frac{dt}{d\rho}$ and is called the Jacobian of w. J be found by noting that

$$\frac{d\,\text{erf}^{-1}(z)}{dz} = \frac{1}{2}\sqrt{\pi}\exp\left[\left(\text{erf}^{-1}(z)\right)^2\right], \quad (6)$$

which gives us, after differentiating using the chain rule,

$$J = \frac{dt}{d\rho} = -\sqrt{2\pi}\exp\left[\left(\text{erf}^{-1}(1-2\rho)\right)^2\right]. \quad (7)$$

Thus, since t has the distribution

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\left[-\frac{1}{2}\left[\frac{t-\mu_t}{\sigma_t}\right]^2\right] \quad (8)$$

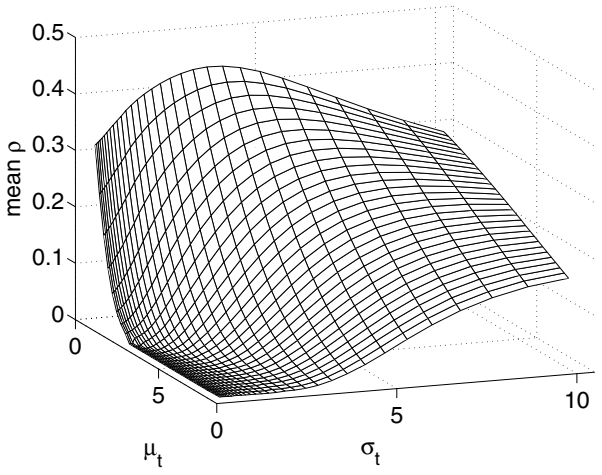

Figure 1. Mean target speaker position for varying μ_t and σ_t.

this leads to the following expression for $g(\rho)$,

$$g(\rho) = \frac{1}{\sigma_t}\exp\left[-\frac{1}{2}\left(\frac{w(\rho)-\mu_t}{\sigma_t}\right)^2 + \left(\text{erf}^{-1}(1-2\rho)\right)^2\right] \quad (9)$$

where $w(\rho)$ is given by equation 4.

6 Performance measures

6.1 Mean target speaker position

The mean target speaker position is given by

$$\bar{\rho} = \int_\infty^\infty \rho g(\rho) d\rho, \quad (10)$$

where $g(\rho)$ is given by equation 9. $\bar{\rho}$ gives the expected proportion of the false speakers who will score higher than the target speaker. This statistic is useful for such purposes as forecasting the average resource utilization per target speaker detection. Figure 1 shows a plot of mean target speaker position for a range of different values of μ_t and σ_t.

6.2 Median target speaker position

In general, as μ_t increases and σ_t decreases, $g(\rho)$ becomes increasingly skewed. This means that usually the target speaker is very close to the top of the queue, but occasionally appears far down the list. In these situations it is unlikely that analysts would continue to examine each sample that far down the queue, especially since in some

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

applications the target speaker may not actually be present in the set under examination. In these cases, a more useful statistic is the median target speaker position $\tilde{\rho}$. This gives the number of entries in the queue that should be examined to successfully locate the target speaker half the time. It can be found from the value of ρ for which the cumulative density function $h(\rho) = 0.5$, where

$$h(\rho) = \int_{\infty}^{\rho} g(s)ds. \tag{11}$$

6.3 Truncated-queue mean target speaker position

A truncated queue of length N is simply the top-N members of a queue, which may or may not contain the target speaker. Let L_q denote the length of the truncated queue which will, with probability q, contain the target speaker. Mathematically,

$$q = \int_{0}^{L_q} g(\rho)\, d\rho. \tag{12}$$

If the experts using the surveillance system examine only the top L_q entries, then the probability of missing the target speaker entirely is $p(m) = 1 - q$. Note that that $\tilde{\rho} = L_{0.5}$. The mean target speaker position for those truncated queues which do contain the target speaker is given by

$$\bar{\rho}_q = \int_{0}^{L_q} \rho\, g(\rho)\, d\rho. \tag{13}$$

$\bar{\rho}_q$ is the average number of false alarms that must be scanned before hitting the target speaker, considering only the truncated queues of length L_q which contain the target speaker. Since this occurs with probability q, it follows that with probability $p(m) = 1 - q$ we must scan L_q false alarms and then fail to detect the target speaker. The number of false alarms scanned per target speaker detected when operating in prioritisation mode is then

$$F_s = \frac{q\bar{\rho}_q + (1-q)L_q}{q}, \tag{14}$$

in comparison to the number of false alarms per target speaker detected in verification mode, which is

$$F_v = \frac{p(f)}{1-p(m)} = \frac{p(f)}{q}. \tag{15}$$

It is apparent from the above equations that the term $q\bar{\rho}_q + (1-q)L_q$ represents the probability that a non-target speaker will result in a false alarm. This allows us to directly compare the performance of a system operating in prioritization mode with the performance of the same system operating in verification mode across a range of values for $p(m)$. For verification mode, $p(m)$ is determined by the system operating point. For prioritization mode, $p(m)$ is determined by the truncated queue length L_q. It is convenient to plot the results on a graph with the same format as the standard DET-plot, and graph false alarms per target speaker detection against the probability of missing the target speaker. However, it is important to remember that the results are not directly comparable with other DET-plots as it is relevant to the prioritization performance only, which incorporates a number of additional mechanisms and assumptions.

7 Modelling example

Let us now re-examine the system of 3.2 with the additional assumptions listed in section 5. First, we decide what values of μ_t and σ_t to use. Real system implementers, of course, are not able to simply choose these parameters, but in this case we are interested in seeing what effect they have on the performance of the system. Obviously we cannot choose purely arbitrary values for both μ_t and σ_t - we must choose such that the operating point $(p(m), p(f))$ lies on the DET curve of the system. First we note that the threshold t must be fixed, since we know that $p(f) = 0.01$ and the false score has a standard normal distributon. Tables of Z-values for standard normal distributions allow us to find that $t = 2.33$. Since the true speaker scores are also normally distributed, we are able to select interesting values of σ_t and find the corresponding μ_t score as

$$\mu_t = t - \sigma z, \tag{16}$$

using the value of z for which $p(x > z | X \sim N(0,1)) = 1 - p(m)$.

We have chosen three distributions somewhat arbitrarily in order to examine the effect on performance. The choice of σ_t was made to ensure a range of different DET curves. The resulting values of μ appear in Table 1.

Table 1. *Parameters for the distributions T_n.*

n	σ_{tn}	μ_{tn}	$\bar{\rho}_n$	$\tilde{\rho}_n$
1	0.2	2.94	0.001964	0.0016
2	1	5.42	0.000065	0
3	5	17.78	0.000246	0

The DET curves of all three systems are shown in Figure 2. It can be seen that all systems intersect at the point $(p(f), p(m)) = (0.01, 0.001)$. Thus, despite their quite different characteristics, all three systems would exhibit the same performance when used in verification mode at that operating point. The mean and median target speaker ranks are given in Table 1.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

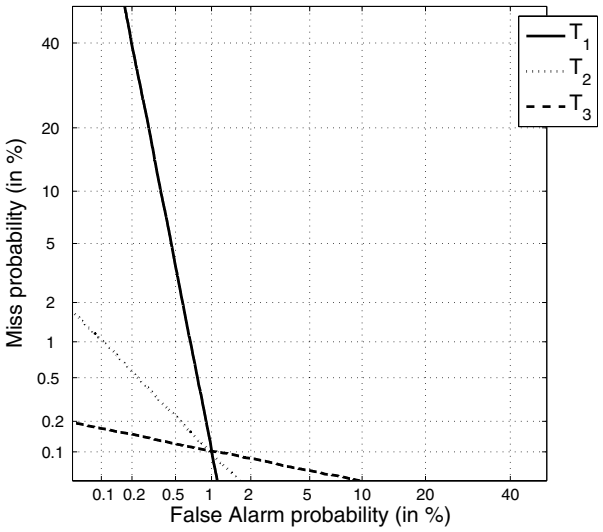

Figure 2. DET curves for the true speaker distributions T_1, T_2 and T_3.

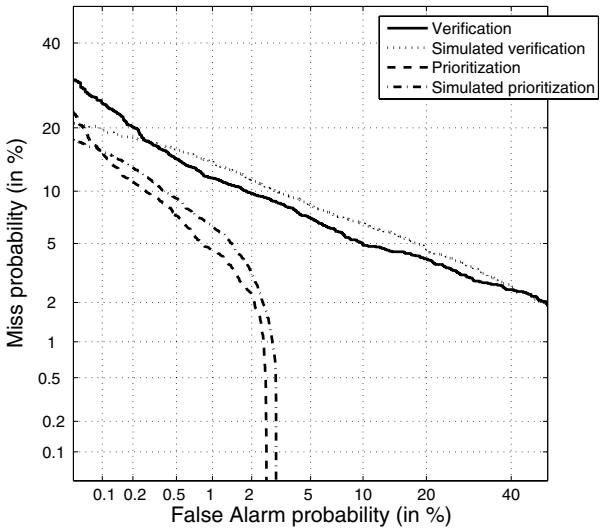

Figure 3. Verification and prioritization performance on NIST SRE 2005 development data, and on simulated score data with the same 1st order statistics.

Using our original example of $N = 10^{11}$, and using the figure for mean target speaker position, we find that the first system gives 196,400,000 false alarms per target, the second system gives 6,500,000 false alarms per target, and the third system gives 24,600,000 false alarms per target on average. Our claim is not that these figures represent a useful surveillance system; rather, it is that these figures represent an improvement of approximately 5 times, 150 times and 40 times respectively over the performance of a system operating in verification mode which makes it clear that surveillance systems are more usefully operated in prioritization mode rather than verification mode.

The median values of these systems are even more interesting. The values of $\tilde{\rho}_n$ listed in Table 1 show that if the systems are to be successful half the time, an analyst need look no further than 16 entries on the queue for system T_1, and only the top entry for systems T_2 and T_3. These figures were produced by Monte Carlo simulations on a queue of 10,000 entries, repeated 100,000 times each, so it is likely that the 'top of the queue' figures for systems T_2 and T_3 would be degraded slightly by simulating using a longer queue. Unfortunately, attempts to calculate mean and median directly from the distribution through numerical integration proved difficult due to the extremely skewed distributions.

8 Experimental comparison

8.1 Data and system description

Experimental demonstration of the queuing approach discussed in this paper was performed on the development data for the NIST Speaker Recognition Evaluation task from 2005. There were 2751 true speaker trials and 28492 false trials. The speaker recognition system was an acoustic GMM-based system. This system has the DET curve as displayed in Figure 3. The outer curve represents the verification performance of the system. The inner curve represents the prioritization performance of the system.

The data was standardized so that $\mu_f = 0, \sigma_f = 1$. This caused the true speaker score distribution to be nearly gaussian with $\mu_t = 5.82, \sigma_t = 2.92$.

8.2 Surveillance performance

2438 of the 2751 true speaker samples were ranked at the top of the queue, and $\bar{\rho} = 0.0255$ and $\tilde{\rho} = 0.00024$. A plot of the prioritization performance is given in Figure 3. It can be seen that the curves representing prioritization bends downward to meet the point $(\bar{\rho}, 0)$ as expected from the definition of $\bar{\rho}$. This results in the prioritization system having dramatically better false alarm rates for the same miss rates when compared to a verification system. A plot of results simulated from pure gaussian data with the same first order statistics as the experimental scores is shown in Figure

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

3. As expected, the simulation displays the same behaviour as the experimental data, but there is a significant performance difference caused by the deviation from the experimental data from gaussianity. This indicates that assessing performance accurately will require either large-scale tests or more accurate modelling of score distributions.

9 Discussion

It is important to remember that the significant performance improvements as measured by $\bar{\rho}$ and $\tilde{\rho}$ do not represent a free performance enhancement. They are predicated on a method of use whereby results are passed to human experts in priority order for subsequent assessment and action. However, the need for such experts in most surveillance systems means that the analysis applies in many cases. In addition, the methods presented only apply to a system which can operate on stored data and is able to defer decision making until all data is available. For example, a system may gather data for 24 hours then present the results to the experts in the form of a queue. A system which is required to present item-by-item decisions will not be able to operate in prioritization mode.

It is possible, indeed likely, that experts will not be able to correctly identify the target speaker with absolute accuracy. This may be taken into account in the above analysis quite simply. If an expert misses the target speaker with probability $p_e(m) = x$, then the effective probability of a truncated queue of length L_q missing the target speaker is $p(m) = x(1 - q)$, and the rest follows.

Other researchers have written about a number of techniques used to improve the performance of speaker recognition systems. We note that it is possible that some techniques may change the behaviour of surveillance systems in a similar manner to the change in performance seen between T_1, T_2 and T_3 in section 7, while not having as significant an impact on the performance of verification systems. The effects of techniques such as those discussed below, and new techniques trialled in the future, should be examined for their effect on queue prioritization performance as described in this paper.

It was noted in [10] that the use of Bayes factor scoring tended to cause a counter-clockwise rotation of the DET curve, which would normally be consistent with an decrease in the σ-ratio. However, it was noted that direct measurement of the variance of the actual true-speaker and false-speaker score distributions revealed that the σ-ratio had increased. The counter-clockwise rotation of the DET curve was interpreted as an effect of the increased Gaussianity (as revealed by negentropy statistics) of the score distributions. Whether Bayes factor scoring is indeed beneficial to surveillance applications is unknown and should be determined by further experiment.

It has been shown that the use of T-norm with large cohort sizes [2], and feature mapping with T-norm [7], also causes a noticeable counter-clockwise rotation of the DET curve. Until further analysis of the cause of this rotation is available, it is unknown whether these techniques are beneficial to surveillance filtering performance.

10 Conclusions

Most surveillance systems need some level of human involvement in the outcomes. As such it may not be necessary to force the automated components of the system to make hard accept/reject decisions. By focussing instead on prioritization of samples, automated systems are able to achieve significant improvements in performance, as measured by false alarms per target detection for truncated queues. A model for the performance of systems having Gaussian output score distributions for target speakers and false speakers was given. This model was shown to have good correspondence to the performance of a real speaker recognition system.

11 Acknowledgment

This work was supported by an Australian Research Council (ARC) Discovery grant No: DP0453278.

References

[1] A. Albiol, L. Torres, and E. Delp. The indexing of persons in news sequences using audio-visual data. *ICASSP*, III:137–140, 2003.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalisation for text-independent speaker verification systems. *Digital Signal Processing*, 10:42–54, 2000.

[3] P. Barger and S. Sridharan. Detection cost bounds for speaker recognition using game theory. *Speaker Odyssey*, 2006 (to appear).

[4] J. Bonastre, F. Bimbot, L. Boe, J. Campbell, D. Reynolds, and I. Magrin-Chagnolleau. Person authentication by voice: a need for caution. *EuroSpeech*, pages 33–36, 2003.

[5] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. *EuroSpeech*, pages 1895–1898, 1997.

[6] J. Navratil and G. Ramaswamy. The awe and mystery of t-norm. *EuroSpeech*, pages 2009–2012, 2003.

[7] D. Reynolds. Channel robust speaker verification via feature mapping. *ICASSP*, II:53–56, 2003.

[8] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[9] B. Schneier. Data mining for terrorists. *Crypto-Gram*, March 15, 2006.

[10] R. Vogt and S. Sridharan. Bayes factor scoring of GMMs for speaker verification. *Speaker Odyssey*, pages 173–178, 2004.