ANALYSIS OF TOP-*K* STRATEGIES FOR OPEN-SET SPEAKER IDENTIFICATION APPLICATIONS

David Colella and Fred Goodman
The MITRE Corporation
March 2012

Abstract

Recent performance gains in speaker verification systems suggest it is now viable to employ these systems in open-set speaker identification applications where an automated decision is passed to a human-in-the-loop for final analysis and decision. This paper examines the performance for when a speaker verification system expands into the identification domain. Our results indicate that separate thresholds should be adopted for the verification and the speaker identification phases. Furthermore, adopting a "top-*k*" approach where the best *k* matches are passed to the analyst for final matching does not greatly improve system detection performance and has a significant impact on overall human workload.

1 Introduction

The purpose of this article is to examine recognition performance and the resulting impact on analyst workload for automated speaker identification scenarios that involve matching an unknown speaker with known speaker models. Voice biometric testing has evolved over time, starting with simple closed-set problems in which the recognition algorithm had to decide which of N speakers were actually speaking. At first, speakers were fully cooperative in providing speech samples, e.g., by voicing strings of numbers, thereby creating text-dependent testing. Commercial applications still use such data. However, in the last decade or so as the performance of speaker identification systems improved, more complex scenarios became possible. The research community now tests with non-cooperative speakers (termed *naïve* speakers) who do not limit their vocabulary or style in any way. This is known as text-independent testing. Although the research community still uses a verification paradigm, i.e., one in which an identify is claimed and a simple yes/no response from the recognition system is required, detection scores are improving to the point where these comparisons can be utilized to identify a given speaker or speakers from a group of known speakers in a crowd. This scenario is said to be an open-set condition, where none, some, or all of the incoming voices may be from known speakers. The analysis of such applications is the focus of this paper. The primary goal is to illustrate in general terms the consequences of particular detection and classification threshold settings on performance and analyst workload for speaker identification scenarios.

Although we focus on speaker identification, the approach presented here is more broadly applicable to other biometrics scenarios that utilize analyst assistance, e.g., for face or iris recognition. As with speaker identification, these other biometrics are likewise faced with similar identity recognition applications. The effectiveness of a recognition system can be measured in various ways and in part depends on the particular objective of an analyst who reviews or is charged with making a final determination from the output of these systems. We will focus on several standard metrics, namely, the probabilities of detection and false accept, the probability of correct classification (i.e., recognition), as well as a metric intended to indicate analyst workload (discussed below).

Our analysis shows that system threshold settings readily adopted and useful for single match detection (verification) are not optimal for detection matching from among a collection of speakers (recognition). The resulting performance in the latter scenario illustrates the importance for the single-match system to operate at an extremely low error rate. Furthermore, the analysis also shows that attempts to employ a "top-k" strategy where not a single ($k = 1$) but the k best matches ($k > 1$) are provided to an analyst for review has a significant impact on the workload burden of the analyst while not providing significantly improved recognition performance. Our results also suggest that alternative thresholds are needed for the single-match and multi-match operations.

This paper is organized as follows. The next section reviews the speaker identification

problem and provides the basic background for the analysis. Section 3 provides the simplifying assumptions used in the analysis and Section 4 discusses detection performance using these assumptions for the single-match speaker verification task. Section 5 introduces the analysis for speaker identification for identifying a given speaker from a larger collection of speakers and Section 6 extends this performance analysis to examine analyst workload issues. In Section 7 we provide our conclusions.

2 Speaker Identification

Text-independent (non-cooperative) automatic speaker identification has been dominated for thirty years by the telephone application [12, 7, 8]. For this application, the question has been whether a given sample of telephone speech is from a known target speaker when a model based on previous speech from the target speaker is already available. Many techniques have been developed for speaker identification (SID) that utilize short-term spectral data, prosody (i.e., pitch, energy, and rhythmic patterns), phoneme sequences, and even word sequences. The National Institute of Standards and Technology (NIST) has historically assumed the role for unbiased competitive testing of the various speaker identification algorithms [15]. The most successful systems in recent evaluations typically fuse at least four subsystems (a subsystem being a component that focuses on one of the above-mentioned characteristics) to arrive at a final yes or no decision. As in all detection problems, there is a trade-off between false accepts and missed detections. In scenarios of interest, automated detections are passed along to an analyst who then listens to the speech and makes a final determination as to whether the identity of the person is as claimed. The trade-offs are commonly weighed in favor of reducing the false accepts since in many practical problems of interest, the probability of a target speaker actually appearing may be quite low. A large number of false accepts requires an analyst to experience the time-consuming frustration of listening to a lot of worthless data. The result is a typical dilemma: decrease the number of false accepts and risk the possibility of a crucial missed detection or face increasing frustration of the analysts having to manually process a large volume of speech data from the overabundance of alerts from the automatic device. Issues regarding system performance and analyst workload associated with how this dilemma is resolved will now be explored.

In recent years, we began considering a different application for SID: the Doctor's office scenario. This is a situation in which a patient visits his Doctor, and the two people sit down and discuss the patient's condition. In this situation, of course, the Doctor asks the patient his name. We record the conversation and attempt to determine whether the patient actually is who he says he is. If the patient is lying about his identity, we then seek the true identity of the person and ask whether this is a person we have previously encountered. Thus, we test the incoming speech against all of the speaker models that are available (and relevant).

Our investigation considers the following scenario (see figure 1). An unknown speaker is presented and claims an identity as one from a list of approved enrolled individuals

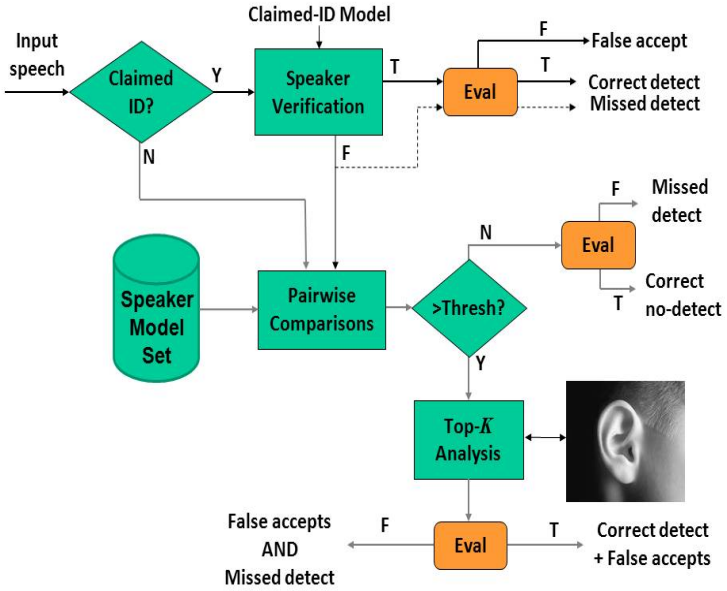

Figure 1: Outline of the overall decision flow for a speaker identification process.

and for which representative models have been developed based on previous samples of speech. For brevity's sake, we will refer to the claimed identity as "*Fred*". Based on a new speech sample collected at the time of claimed identity, a decision is made as to whether or not the claimed identity *Fred* is correct. This is accomplished by matching the speech sample with the model of *Fred*. We consider this portion of the scenario as the verification phase. However, in our discussion we will most often use the term *detection* to mean the process of verifying a speaker who is claiming a particular identity. The *probability of detection* P_d is the probability that if the claimed identity is the correct one, then we are able to match the new speech with the stored model. A *missed detection* results if the new speech is not an adequate match for the claimed (and true) identity, resulting in a failure to properly recognize the speaker. If the claimed identity is a false one and our decision allows the claimed identity, then this is a *false accept* and the probability P_{fa} that this occurs is referred to as the *probability of false accept*. Additional details of the specific methods for this decision process are provided in the following sections.

The situation in which our initial decision is to reject the claimed identity as being true (regardless of whether this is accurate or not) is referred to as the classification phase. During this phase we are concerned with either the case where the speaker is an impostor (i.e., not *Fred*) and correctly recognized as not *Fred* or the missed detection case where *Fred* actually presents himself but is not recognized. We then need to evaluate the speech segment for matching against the other speech models from a collection of N individuals to see if we can correctly identify the impostor.

If the impostor happens to be one of the speakers from the collection of known speakers and we can correctly recognize his identity or if he is not one of the known speakers and

we correctly recognize that he is not, then we refer to this as a *correct classification* and the probability P_c of performing this correctly is the *probability of correct classification*. Otherwise, we have a *mis-classification* whose probability is $P_m = 1 - P_c$. One of the considerations we will explore is the effect of using the less restrictive condition where it is sufficient to declare a correct classification of the impostor when the speaker is one of the top k matches, $N \gg k > 1$, and not require the correct model to be the best match (i.e., $k = 1$).

3 Simplifying Assumptions

In order to keep our effort tractable, we make a number of simplifying assumptions. These assumptions, although not reflecting exact operating conditions, remain reasonably realistic for most cases. Furthermore, they are sufficient to highlight the relationship between detection and classification characteristics and analyst effectiveness. We begin by assuming we have a collection of $N+1$ speakers as well as sufficient samples of speech for each speaker to build a model M_n that represents *recognizable* speech patterns for speaker n; e.g., see [4].

When presented with an unknown speaker, a test of that speaker against the model M_n produces an output statistic x that can be used to make a recognition decision that the unknown speaker is or is not speaker n. Generally, x will be determined from the combination of a number of model components, or subsystems, after which a fusion process is applied to determine the single value x, e.g., cf. [3]. We will not delve deeply into this process but it is nonetheless worthwhile to briefly discuss this aspect of speaker identification. In most speaker identification systems, the output statistic x is determined through a log-likelihood ratio, namely, the log-probability that the speaker is *Fred* minus the log-probability that the speaker is not *Fred*. An output result greater than zero means that *Fred* is more likely to be the speaker than an impostor. The detection hypothesis is of course based on the model for *Fred*. The alternative hypothesis that the speaker is *not Fred* is usually based on what is referred to as a Universal Background Model (UBM) [11]. The UBM is constructed using the pooled speech of a large number of unknown speakers. Usually, a separate UBM (or equivalent) is developed for each speech feature type (spectral, prosody, phoneme sequences, etc.).

The statistic x for a single fixed model M_n can be considered a random variable when tested against a (large) collection of unknown speakers. If an unknown speaker is different from the speaker used to build the model, we say that speaker is an *impostor*. It is generally true that the the output x for impostors can be modeled as a gaussian distributed random variable. This distribution, and hence corresponding statistic, is renormalized so that x becomes a zero-mean unit-variance gaussian:

$$\mathbf{E}\left[x \,|\, \text{impostor}\right] = 0 \;\text{ and }\; \mathbf{E}\left[x^2 \,|\, \text{impostor}\right] = 1.$$

It is also generally assumed that the corresponding renormalized output for when the unknown speaker is tested against his actual (i.e., true) model is likewise gaussian, only now

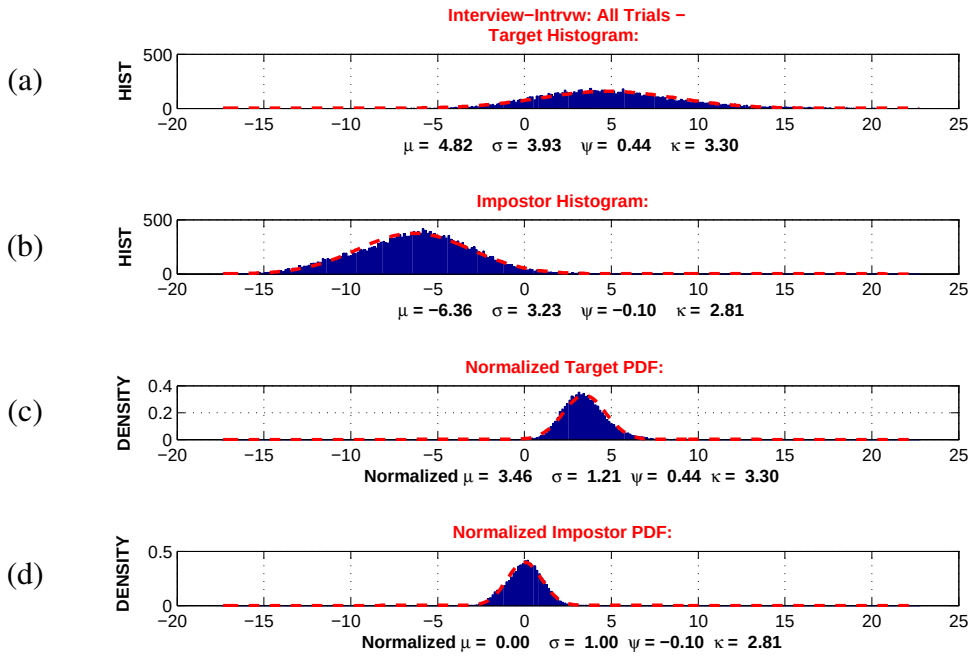

Figure 2: An example illustrating the renormalization of the densities for speaker identification. Plots (a-b) are the original distributions; renormalizing the impostor distribution to have zero mean and unit variance (d) then leads to a true model distribution (c) with mean $\mu > 0$ and variance $\sigma^2 > 1$.

with mean μ_n and variance σ_n^2:

$$\mathbf{E}\left[x \,|\, \text{true model}\right] = \mu_n \;\text{ and }\; \mathbf{E}\left[(x-\mu_n)^2 \,|\, \text{true model}\right] = \sigma_n^2.$$

We indicate the effect of renormalization on the impostor and true model distributions in figure 2. The histogram data is from the NIST SRE-08 evaluation [16]. The data represents more than 11,000 target trials and more than 22,000 impostor trials. The curves shown are typical curves for speaker identification systems.

Our first major assumption is that the renormalized gaussian true model distributions are identically distributed, i.e., that $\mu_n = \mu$ and $\sigma_n^2 = \sigma^2$ for all n. This assumption greatly simplifies many of our later calculations, e.g., equation (12), and keeps our computations more tractable. Significant empirical evidence [9] obtained from the NIST SRE-08 evaluation data also indicates that

$$\sigma > 1.$$

This assumption is somewhat surprising, since it would normally be expected that the variation in speaker output statistic would be more confined for an arbitrary speech sample tested against that speaker's own model. However, as a consequence of the overwhelming results to the contrary we align our assumption with these empirical results. It is also the case that $\mu > 0$.

4 Detection Strategy

We begin with the detection problem. To reiterate, an unknown speaker is presented and claims to be a known (i.e., modeled) person, *Fred*. Our first task is to determine if this is correct. The speech sample from the speaker is analyzed and compared to the model for *Fred*, resulting in a test statistic x. Thus, when the unknown speaker is not *Fred*, the density function $p_{F,0}(x)$ for the statistic becomes:

$$p_{F,0}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The above expression assumes that we have normalized the parameter x so that $p_{F,0}$ will be zero mean with unit variance. There are a variety of ways to make a decision based on x. We initially will assume *maximum likelihood* (ML) detection. Later we will also consider the equal error rate (EER) as this is a common metric for speaker identification applications. In addition we require the density function that measures the variation of the output of *Fred*'s model when *Fred* is the actual speaker, and need to normalize these outputs in a way that is consistent with the normalization for $p_{F,0}$. Under our assumptions, this density can therefore be expressed as

$$p_{F,1}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

We need to establish a threshold whereby if the sample x is more representative of the density $p_{F,1}$ than $p_{F,0}$ then we accept the hypothesis that the speaker is indeed *Fred*. For ML detection, this threshold is established by the condition

$$\text{ACCEPT } \textit{Fred} \text{ when } \; p_{F,1}(x) \geq p_{F,0}(x). \tag{1}$$

This assumes of course that an impostor claiming to be *Fred* is equally likely to appear as *Fred* himself. Later we consider the Bayesian modifications for when this is not the case along with applying a cost for making any particular decision. For now, we accept the speaker as *Fred* when and only when a sample x satisfies

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \geq \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

that is, when

$$\frac{1}{\sigma} e^{-(x-\mu)^2/2\sigma^2 + x^2/2} \geq 1.$$

Taking logarithms and simplifying, we arrive at the condition

$$\left(x + \frac{\mu}{\alpha^2}\right)^2 \geq \eta_1^2 \tag{2}$$

where $\alpha^2 = \sigma^2 - 1$, $\alpha > 0$, and

$$\eta_1^2 = \frac{\sigma^2}{\alpha^2} \log \sigma^2 + \frac{\sigma^2}{\alpha^4} \mu^2, \quad \eta_1 > 0. \tag{3}$$

Note that since $\sigma > 1$ the expression on the right-hand side of equation (3) will be positive. The *two-sided* test in equation (2) says that the unknown speaker should be accepted as *Fred* if and only if the sample statistic x satisfies this constraint. One interpretation of this is that the test accepts *Fred* only by deciding that the unknown speaker is *not* an impostor!

More generally, the development above could include the a priori probability ρ_1 (respectively $\rho_0 = 1 - \rho_1$) that the unknown speaker who claims to be *Fred* is (respectively is not) *Fred*, as well as the cost functions $c(n,n'), n,n' = 0,1$, associated with choosing not *Fred* ($n = 0$) or *Fred* ($n = 1$) when in fact the speaker either is not *Fred* ($n' = 0$) or is *Fred* ($n' = 1$). The goal is then to determine an appropriate threshold so that a cost function, e.g.,

$$\begin{aligned} C &= \rho_0 \left[c(0,0)\ \text{Prob}\{\text{choose not } \textit{Fred} | \text{not } \textit{Fred}\}\right. \\ &\qquad \left. + c(1,0)\ \text{Prob}\{\text{choose } \textit{Fred} | \text{not } \textit{Fred}\}\right] \\ &\quad + \rho_1 \left[c(0,1)\ \text{Prob}\{\text{choose not } \textit{Fred} | \textit{Fred}\}\right. \\ &\qquad \left. + c(1,1)\ \text{Prob}\{\text{choose } \textit{Fred} | \textit{Fred}\}\right] \end{aligned}$$

is minimized. For this general result we now choose *Fred* when the test statistic x satisfies

$$\frac{p_{F,1}(x)}{p_{F,0}(x)} \geq \frac{\rho_0}{\rho_1} \cdot \frac{c(1,0) - c(0,0)}{c(0,1) - c(1,1)}. \tag{4}$$

If all of the cost functions for an incorrect decision are equal and those for making a correct decision are zero, then the quotient involving those terms disappear. This was the case for our previous discussion, and along with the assumption that $\rho_0 = \rho_1 = 1/2$ we see that (5) reduces to our previous test. For the detection problem, it would not be unreasonable to assume that the cost associated with making a correct decision is zero, i.e., $c(0,0) = c(1,1) = 0$. The cost of a missed detection $C_{md} = c(0,1)$ is associated with increased (unnecessary) workload since we would now need to check other speakers from our collection and could require analyst intervention; the cost $C_{fa} = c(1,0)$ is associated with a mis-identification of an impostor, i.e., a false accept. Depending on the scenario (the availability of human labor, the value of detecting *Fred*, etc.) the costs could vary widely. In the NIST evaluations the values $C_{md} = 10$ and $C_{fa} = 1$ have been used, representing a situation where detecting *Fred* is quite important.

We can now write our threshold condition in terms of the a priori probabilities ρ_0, ρ_1 and costs C_{md}, C_{fa}. The condition is:

$$\left(x + \frac{\mu}{\alpha^2}\right)^2 \geq \eta^2 \tag{5}$$

where now

$$\eta^2 = \eta_1^2 + \frac{\sigma^2}{\alpha^2} \log\left(\frac{\rho_0 C_{fa}}{\rho_1 C_{md}}\right)^2. \tag{6}$$

We still require $\eta^2 > 0$, so that if the quotient in the log term on the right-hand side is too small, the entire expression becomes negative and no such threshold exists that will

minimize cost. (Technically, the cost is then minimized for a threshold equal to 0.) As the cost for a false accept increases, this translates into making the threshold larger, i.e., becoming more restrictive for deciding that the unknown speaker is not an impostor.

As one final note, we point out that if in fact $\sigma^2 = 1$ our two-sided test reduces to a one-sided test on the test statistic x:

$$\text{choose } \textit{Fred} \text{ if } \; x \geq \mu/2.$$

The probability of detection can now be computed. From our discussion above, this is:

$$P_d = \int_{R_\eta} p_{F,1}(x)\, dx \tag{7}$$

where R_η is the region defined by $\{x : |x + \mu/\alpha^2| \geq \eta\}$. The probability of a missed detection becomes $1 - P_d$. Similarly, using the same threshold condition η, the probability of false accept is obtained using the density $p_{F,0}$:

$$P_{fa} = \int_{R_\eta} p_{F,0}(x)\, dx \tag{8}$$

since this expresses the likelihood that the statistic for an impostor claiming to be *Fred* will lie within the accept region.

An often-used alternative to maximum likelihood detection is the equal error rate (EER) condition. Taking this approach is more common in speaker identification systems (e.g., cf. [6]). For the equal error rate, the test (5) remains the same except now the threshold used is not determined from the condition (1) or (6) but instead η is chosen so that the missed detection and false accept probabilities are equal:

$$P_{md} = 1 - P_d = P_{fa}. \tag{9}$$

We illustrate detection performance for these detection strategies with several detection curves in figures 3 and 4. Generally, the detection statistics, probability of false accept and probability of missed detection, will depend on the two parameters μ and σ. Our graphs illustrate these statistics for relatively small values of those parameters to highlight the general trends. As expected, the error rates decrease for increasing mean and increase for increasing variance (figure 3). This also provides some guidance as to the needed values of those parameters if a certain performance is desired. These curves highlight the independent functional dependence of the error rates on the mean and variance. Close inspection not surprisingly indicates that generally the mean must be large compared to the variance in order to achieve reasonable error rates, i.e., error rates not greater than 1 or 2 per cent.

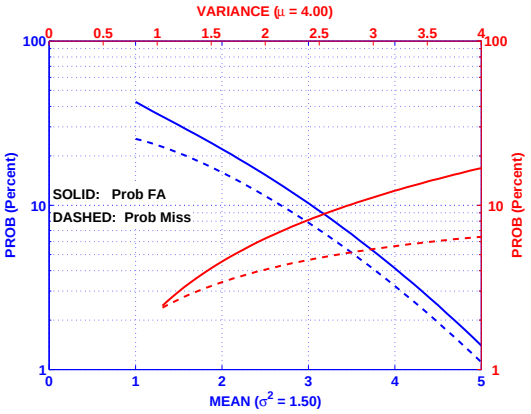

Figure 3: Probability of false accept (solid lines) and missed detection (dashed lines) for ML detection. Blue curves vary the mean μ for $\sigma^2 = 1.5$ while the red curves vary σ^2 for fixed mean $\mu = 4.0$.

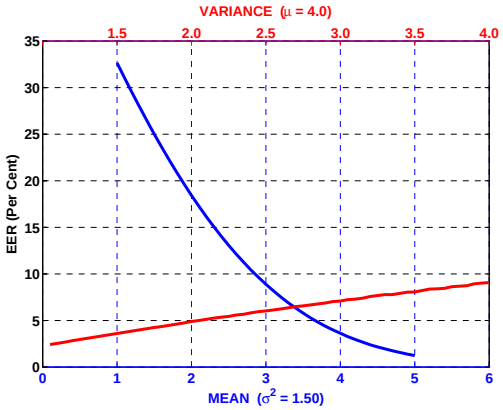

Figure 4: The equal error rate for different means (μ) for a fixed variance ($\sigma^2 = 1.5$) in blue and also for a fixed mean ($\mu = 4.0$) with different variances (σ^2) in red.

In figure 4 we provide a plot of the equal error rates (EER) for the same values of the parameters used in figure 3. In this case, the threshold is determined by equalizing the false accept and missed detection probabilities. The trends suggest that the EER is more sensitive to the mean than to the variance. They indicate relative constraints on our parameters μ and σ^2 in order to maintain even a 10% EER. Together, the plots in figures 3 and 4 provide a guideline for how well a system must separate (or distinguish) the impostor speakers from a given speaker in order to attain a particular level of system performance.

Although both the EER and maximum likelihood detection strategies can be used to set verification thresholds, figure 5 compares the ML false accept and missed detect probabilities to the EER for EER rates up to 10% as determined by fixing $\sigma^2 = 1.5$ and varying the mean. As this plot indicates, for EER probabilities less than about 2 or 3 percent there is little difference between the EER errors and the maximum likelihood errors. Of course, this comparison will be altered for different variances and means since the mapping between σ^2, μ and EER is not one-to-one. (For comparison, if $\sigma^2 = 1.0$ then the ML and EER probabilities will be equal.) In an identification system where our known collection of speakers is very large ($N \gg 1$) these small differences can result in significant variations in recognition performance. One purpose of this paper is to highlight the impact these variations could have for our recognition task, regardless of which detection strategy is applied. For simplicity we will therefore primarily use the EER criterion as a baseline and present further results based on this condition.

Finally, we illustrate the effect of incorporating a priori probabilities as well as cost functions for maximum likelihood detection strategies. In addition to the equal priors and

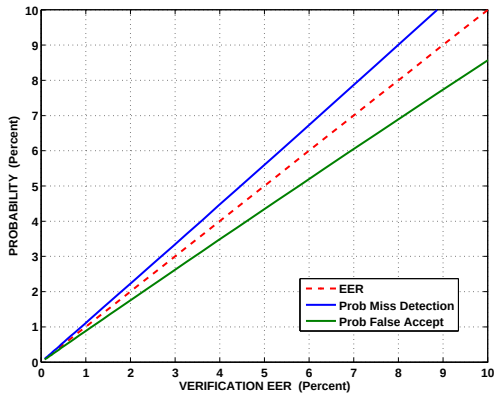

Figure 5: The maximum likelihood probabilities of missed detection and false accept as a function of the equal error rate when determined by $\sigma^2 = 1.5$ and varying μ.

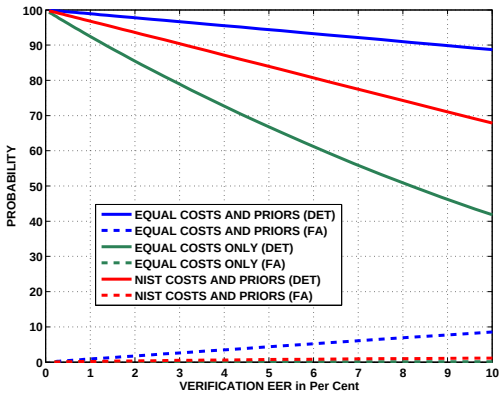

Figure 6: Detect (solid) and false accept (dashed) performance using three ML thresholds: equal costs and priors (blue), equal costs and $\rho_1 = 0.01$ (green), and the NIST parameters (red).

equal costs threshold employed for most of the results of this paper (equation (3)), here we also consider the cases when the costs are equal with the a priori probabilities set at $\rho_1 = 0.01, \rho_0 = 0.99$ and when the cost and a priori probabilities are set with the NIST evaluation values $\rho_1 = 0.01, \rho_0 = 0.99, C_{md} = 0.91, C_{fa} = 0.09$. Detection performance for these strategies is shown in figure 6. This plot illustrates the fact that a strategy including cost and a priori parameters does not always lead to maximizing detections (solid lines). The reason for this is that the relatively high prior probability for ρ_0 puts a significant emphasis on reducing the false accept rate. This reduction is shown as the dashed lines in the figure, where the false accept probabilities for the strategies using $\rho_0 = 0.99$ are a fraction of that for the case of equal priors.

5 False Identities

We now address the issue of when an impostor is correctly identified as not being *Fred*. Our interest then becomes whether we can correctly recognize this speaker as a member from a larger set of speakers. This entails computing a collection of statistics $x_n, n = 1, \ldots, N$ using each of the N remaining models. For a fixed integer $k \geq 1$, a decision is then based on the best k comparisons that also exceed the given threshold condition. Keep in mind that the output statistics x_n are the result of comparisons between each speaker model and the UBM. This approach makes the question of thresholding much simpler than if we compared the speaker model likelihoods to each other. Instead, we perform N speaker detections in parallel. In effect, the UBM comparison acts as a normalizing process. The result is that the models with the k largest margins over the UBM will be presented to the analyst for the final determination.

Let us now recall the ML condition expressed in equation (5). This test suggests that the decision to find the best model fit from the larger set of speakers can be accomplished by taking the k models for which the squared term on the left-hand side in (5) is largest (and provided that the threshold condition (5) is met). Using the squared expression simplifies our test somewhat and allows us to consider order statistics to examine our detection and classification problems. Since for each model the output statistic x is a gaussian distributed random variable, the new variable

$$y = \left(x + \frac{\mu}{\alpha^2}\right)^2 \tag{10}$$

is essentially a (noncentral) χ^2 distribution with one degree of freedom, e.g., [10]. The random variable y necessarily satisfies $y \geq 0$. The threshold test from equation (5) for y becomes:

$$y \geq \eta^2. \tag{11}$$

Now, the order statistics of a sequence of non-negative random samples $\{y_n\}$ is a re-sorting of the variables in increasing order. Specifically, given samples $\{y_n\}$, the k-th order statistic $y_{(k)}$ is the k-th smallest element. In other words, if we order the samples from smallest to largest

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(N)}$$

where the parentheses in the subscripts indicate the reordering index, then the k-th order statistic is the k-th element $y_{(k)}$ in this list. Since the original elements y_n are random variables, the order statistics are likewise random variables. Furthermore, expressions for the density function of these statistics are known. In a case like ours where the random variables are independent and identically distributed, if $f(y)$ is the density function for the y_n and $F(u) = \int_{-\infty}^{u} f(y)\,dy$ is its cumulative distribution function (CDF), then the density function for the k-th order statistic $y_{(k)}$ is (cf. [5]):

$$p_{(k)}(y) = \frac{N!}{(k-1)!(N-k)!} F(y)^{k-1} \left(1 - F(y)\right)^{N-k} f(y) \tag{12}$$

where of course $n!$ represents the factorial product $1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$.

In order to determine the associated probabilities for classification, we first look more closely at the random variable y. In all cases, y is non-central χ^2 with non-centrality parameter λ that depends on the model being tested. Since $x + \mu/\alpha^2$ has mean μ/α^2 and variance one when the model being tested is not that of the unknown speaker then the non-centrality parameter is

$$\lambda_0 = \frac{\mu^2}{\alpha^4} \quad \text{for impostor.} \tag{13}$$

The density function of y is then given by [10]:

$$p_0(y) = \frac{1}{2} e^{-(y+\lambda_0)/2} \left(\frac{y}{\lambda_0}\right)^{-1/4} I_{-1/2}\left(\sqrt{\lambda_0 y}\right), \quad y \geq 0 \tag{14}$$

where $I_{-1/2}(y)$ is a modified Bessel function of the first kind of order $-1/2$ and most easily expressed as the power series [2]:

$$I_{-1/2}(y) = \left(\frac{y}{2}\right)^{-1/2} \sum_{r=0}^{\infty} \frac{1}{r!\,\Gamma\left(r+\frac{1}{2}\right)} \left(\frac{y}{2}\right)^{2r} \tag{15}$$

and Γ is the standard Gamma function $\Gamma(u) = \int_0^\infty t^{u-1}e^{-t}\,dt$. The cumulative distribution function F for p_0 is calculated to be

$$F(y) = \sum_{r=0}^{\infty} e^{-\lambda_0/2} \frac{(\lambda_0/2)^r}{r!} \gamma_{r+1/2}(y/2) \tag{16}$$

where γ_u is the (lower) incomplete gamma function

$$\gamma_u(v) = \frac{1}{\Gamma(u)} \int_0^v t^{u-1}e^{-t}\,dt.$$

(This is easily verified by taking the derivative of F and comparing to equations (14-15).)

The situation where we have a model match is similar but slightly modified. In order to develop the density function for y in this case we first consider the new variable $y_1 = y/\sigma^2 = ((x+\mu/\alpha^2)/\sigma)^2$. Since x now has mean μ and variance σ^2, the variable $(x+\mu/\alpha^2)/\sigma$ has mean $\sigma\mu/\alpha^2$ and variance one so that y_1 is again non-central χ^2 with non-centrality parameter

$$\lambda_1 = \sigma^2\mu^2/\alpha^4 \quad \text{for model match.} \tag{17}$$

As for y previously, the density for y_1 is determined from the expression in equation (14) except with y being replaced by y_1 and λ_0 replaced by λ_1. The change of variable from y_1 back to $y = \sigma^2 y_1$ therefore results in the following density for the case of model match:

$$p_1(y) = \frac{1}{2\sigma^2} e^{-(y/\sigma^2+\lambda_1)/2} \left(\frac{y}{\lambda_1\sigma^2}\right)^{-1/4} I_{-1/2}\left(\sqrt{\lambda_1 y/\sigma^2}\right), \quad y \geq 0. \tag{18}$$

There are now two cases to consider for computing the probability of correct classification P_c. The first, and easier, case is when the impostor is not in our larger set of speakers. A correct classification in this case would entail the output y_n for all models to fail the threshold test:

$$y_n < \eta^2, \quad \text{for all } n.$$

The conditional probability $P_{c,0}$ of correctly recognizing that the speaker is not in the set is obtained by noting that the maximum of $\{y_n\}$, which equals the N-th order statistic $y_{(N)}$, is less than our threshold:

$$y_{(N)} < \eta^2,$$

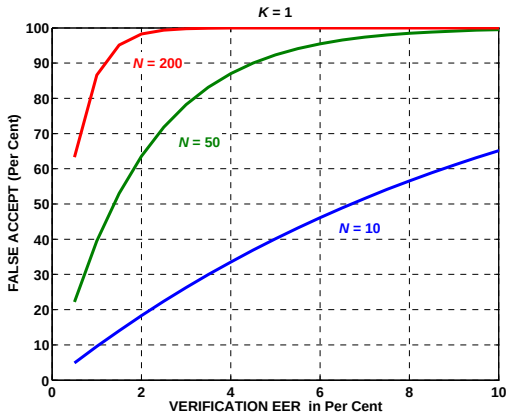

Figure 7: False accept rate for accepting a person not on the speaker set for different set sizes as a function of EER based on $\sigma^2 = 1.5$ and varying μ; results shown use the single best match (i.e., $k = 1$) exceeding threshold.

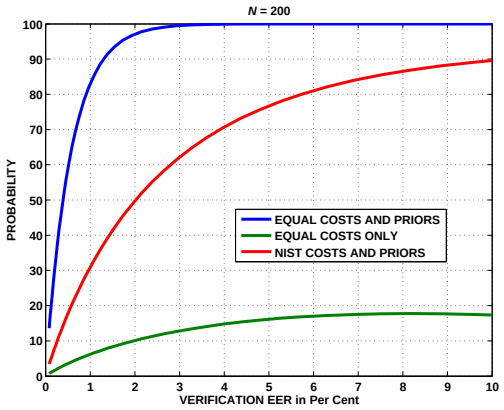

Figure 8: The false accept rate for speaker set size $N = 200$ for different threshold strategies using equal costs and priors (blue), equal costs and $a_1 = 0.01$ (green), and the NIST costs and priors parameters (red).

and given our expression (12) for the density when $k = N$ we have

$$\begin{aligned} P_{c,0} &= \int_0^{\eta^2} NF(y)^{N-1} p_0(y)\, dy \\ &= F(\eta^2)^N \end{aligned} \tag{19}$$

where F is the cumulative distribution function for p_0 as in (16) and p_0 is given by equation (14) with non-centrality parameter given in equation (13). The expression $F(\eta^2) = 1 - P_{fa}$ is the probability for correctly identifying the speaker is not *Fred*. Of course it is readily realized that the above integral is an alternate form for an integral in our original variable x:

$$\begin{aligned} F(y) &= \int_{\{x:(x+\mu/\alpha^2)^2 < y\}} p_{F,0}(x)\, dx \\ &= \int_{R_y} p_{F,0}(x)\, dx \end{aligned} \tag{20}$$

where $R_y = \{x : -\frac{\mu}{\alpha^2} - \sqrt{y} < x < -\frac{\mu}{\alpha^2} + \sqrt{y}\}$. Thus, equations (19) and (20) together give the probability of correct classification for this case.

These results are illustrated in figure 7 for the simplest case of $k = 1$ for EER based on $\sigma^2 = 1.5$ and varying μ. The figure plots the false accept rate for the three cases of $N = 10, 50$, and 200 where the false accept probability is given by $1 - P_{c,0}$. The false accept rates shown here (and in figure 8) are error rates for accepting a speaker for whom we have no speaker model, and does not include the possibility of misclassifying this speaker when

we do have his model (i.e., misidentifying the speaker as some other than himself). As can be seen, even at low EER the false accept rate increases rapidly for the different set sizes. This increase in false accept rate has a significant workload impact by creating a large increase in the number of speech samples that must be passed to the analyst for evaluation. Even in the case when $N = 50$, an EER of 5% leads to more than 90% false accepts. It it clear that the false accept rate is a function of N but not k. This is because a false accept occurs when the threshold is exceeded for even a single test, so that the false accept probability remains constant regardless of k. In comparison, as N increases the likelihood of one of the output statistics exceeding the threshold test also increases. The explicit dependence of the false accept rate on N is expressed in the relation (19).

In figure 8 we show the effect of using different threshold strategies on false accept rates when our set size is $N = 200$. The threshold strategies are the same as those highlighted in figure 6: equal costs and a priori probabilities (blue), equal costs and a priori probabilities $\rho_1 = 0.01, \rho_0 = 0.99$ (green), and the NIST a priori probabilities $\rho_1 = 0.01, \rho_0 = 0.99$ and costs $C_{md} = 0.91$, $C_{fa} = 0.09$ (red). This plot supports the argument that in the case of a relatively small target prior ρ_1, using the a priori probabilities in threshold setting as in (6) helps drive down the overall false accept rate. It suggests that adopting a threshold based on the a priori probabilities could be more beneficial for reducing analyst workload issues. In the next section we'll continue to explore this comparison relative to analyst workload.

Finally, note also that from the expression (19) we can determine limitations on the system. For example, if a threshold is set to give a detection probability $\beta = F(\eta^2)$ and a particular correct impostor recognition probability P_{des} is desired, then the capacity in terms of the number of speakers that we can have in our set is limited by:

$$N \leq \frac{\log P_{\text{des}}}{\log \beta}.$$

Similarly, if $P_{\text{des}} = 0.5$ and $N = 1000$ then we must set our threshold η so that $\beta \geq P_{\text{des}}^{1/N} = 0.9993$. Since $F = 1 - P_{fa}$, this constraint would require our system to maintain a false accept rate of 0.07%. For the best current generation systems (those satisfying $\mu \geq 4, \sigma^2 \leq 1.5$) such a false accept rate would result in a detection rate no better than about 70% while on average half of the impostors would still be accepted for analyst review (since $P_{\text{des}} = 0.5$). Of course, the above doesn't take into account the a priori probability for this speaker to be a known speaker, which could modify these conclusions significantly.

The second case is when the model for a speaker is known. In this case, correct classification requires that the output statistic x_ℓ for the true model ℓ satisfy both the threshold condition $y_\ell = (x_\ell + \mu/\alpha^2)^2 \geq \eta^2$ as well as being one of the k largest values:

$$y_\ell \geq y_{(N-k+1)}.$$

The threshold condition is thus true only for y greater than or equal to the threshold value, and the density for such a value y is given by equation (18) with λ_1 given by equation (17).

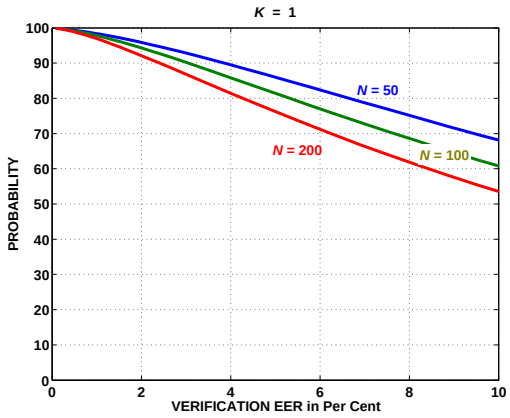

Figure 9: Probability of correctly identifying a speaker from a set of speakers as a function of EER based on $\sigma^2 = 1.5$ and varying μ where only the single best match is considered for analyst evaluation.

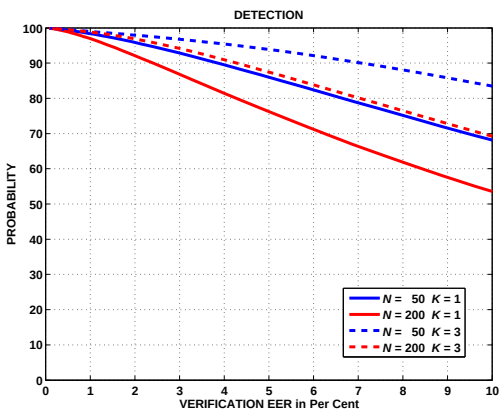

Figure 10: A comparison of the probabilities for correctly detecting a speaker from the set of speakers for several values of N and k expressed as a function of verification EER determined by $\sigma^2 = 1.5$ and varying μ.

For each such fixed value $y = y_\ell$, the conditional probability that $y \geq y_{(N-k+1)}$ is equal to the probability that y is greater than or equal to the $(k-1)$-th largest of the $N-1$ output statistics for the models that remain after the ℓ-th model is removed. We write this as:

$$y \geq \tilde{y}_{(N-k+1)}$$

where $\tilde{y}$ represents the fact that we consider only $N-1$ models. The associated conditional probability is given by:

$$\text{Prob}\{\tilde{y}_{(N-k+1)} \leq y \,|\, y\} = \int_0^y \tilde{p}_{(N-k+1)}(u)\,du = \tilde{F}_{(N-k+1)}(y)$$

where $\tilde{F}_{(n)}$ is the cumulative distribution function for $\tilde{p}_{(n)}$ which is given by equation (12) for the impostor density $f(u) = p_0(u)$ and replacing N in that equation by $N-1$ after accounting for the true model variable term for y_ℓ. Explicitly, $\tilde{p}_{(N-k+1)}$ is given by

$$\tilde{p}_{(N-k+1)}(u) = \frac{(N-1)!}{(k-2)!(N-k)!} F(u)^{N-k} (1-F(u))^{k-2} f(u)$$

with F the CDF of the density $f(u) = p_0(u)$ as in equation (14) with $\lambda_0 = \mu^2/\alpha^4$. Therefore, the probability of correct classification in this second case is

$$\begin{aligned} P_{c,1} &= \int_{\eta^2}^{\infty} \text{Prob}\{\tilde{y}_{(N-k+1)} \leq y \,|\, y\} p_1(y)\,dy \\ &= \int_{\eta^2}^{\infty} \tilde{F}_{(N-k+1)}(y) p_1(y)\,dy. \end{aligned} \tag{21}$$

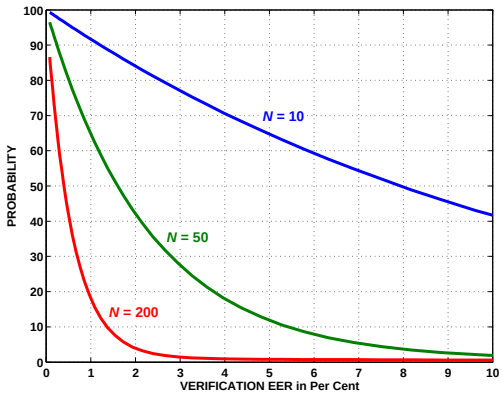

Figure 11: Probability of correct classification for several values of set size N (with $k = 1$) as a function of EER with $\sigma^2 = 1.5$ and varying μ; equal a prior probabilities and costs are assumed.

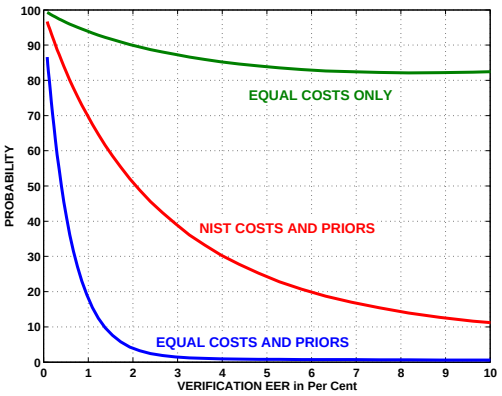

Figure 12: Correct classification for a set size of $N = 200$ using different threshold strategies: equal costs and priors, equal costs and prior $\rho_1 = 0.01$, and the NIST costs and prior parameters.

Figure 9 shows the probability of correctly detecting a speaker from our set of known speaker models for several set sizes. As can be seen, at least for $k = 1$, there is not much difference in detection performance as N increases, particularly for EER at 2% or less. Also, figure 10 shows that as expected detection performance is improved when adopting a strategy for an analyst to consider a larger number of speaker samples. For both set sizes of $N = 50$ (blue curves) and 200 (red curves) an improvement is seen as we increase k from 1 (solid curves) to 3 (dashed curves), indicating that the improvement would increase as N increases. For EER rates close to 2%, the missed detection probability is reduced by as much as 50%, a significant improvement. These plots also suggest that as N increases an appropriate strategy for improving detection is to increase k accordingly. However, as we shall see, this strategy has an undesirable consequence for analyst workload.

Combining equations (19) and (21) we obtain the probability of correct classification as

$$\begin{aligned} P_c &= P_{c,0} \cdot \text{Prob}\{\text{not in the speaker set}\} + P_{c,1} \cdot \text{Prob}\{\text{in the speaker set}\} \\ &= P_{c,0} \cdot \rho_0 + P_{c,1} \cdot \rho_1, \end{aligned}$$

where with a slight abuse of notation we use ρ_0 (respectively ρ_1) to denote the a priori probabilities that the speaker is not (respectively is) in the set of known speaker models. When ρ_1 is small, P_c will be dominated by performance for false acceptance. Figures 11 and 12 highlight this fact. In figure 11 we show the probability of correct classification for several values of set size for known speaker models, $N = 10, 50$, and 200, assuming equal cost functions and a priori probabilities. The curves were of course generated using $k = 1$ since a false alarm occurs if only a single statistic exceeds threshold. The performance curves for using a strategy of $k > 1$ will in general improve performance since the

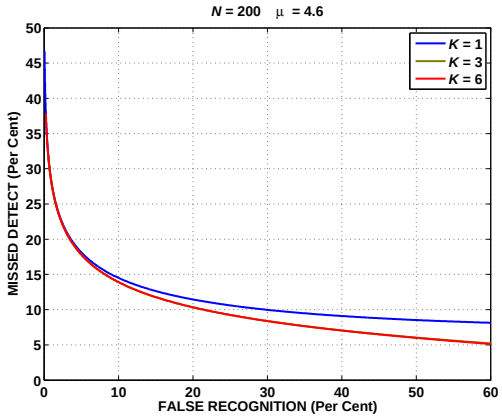

Figure 13: Detection curves measuring system detection probability versus false recognition probability for different values of k using set size $N = 200$, $\sigma^2 = 1.5$, and μ associated with EER = 2%.

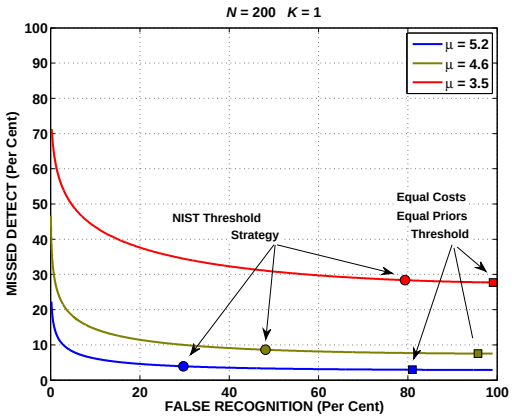

Figure 14: Detection curves measuring system detection probability versus false recognition probability for set size $N = 200$, $\sigma^2 = 1.5$ and different μs associated with EERs equal to $1\%, 2\%$, and 6%; performance for NIST parameter and standard ML equal costs and priors threshold strategies are also indicated.

false accept rate for a speaker whose model is not in our set remains unchanged while the probability for detection will improve as k increases (see figure 10). However, this will be significant only when ρ_1 is not small since this improvement in detection is weighted by this a priori probability. Figure 12 compares the correct classification probability for the three strategies for choosing a threshold discussed earlier. As shown, an improvement in overall classification can be achieved by altering the threshold strategy for detection. In both figures, it should be noted the similarities with figures 7 and 8, where the graphs in the above figures are virtual reciprocals of the graphs in the previous ones. This demonstrates how much overall performance is driven by the false accept performance.

The examination of performance for system classification is not yet complete. For the case when we have a model for the speaker is our set, the above describes the probability associated with correctly detecting the speaker within our set in the sense that the true model statistic exceeds threshold and is one of the top k. An error occurs when either no statistic exceeds threshold or when the speaker is not one of the top k and at least one mismatched model statistic exceeds threshold. Let us now consider the second of these, a sort of false detect case. Making this type of error would require two conditions to hold. First, at least one output statistic for the $N-1$ non-match models must exceed threshold. Thus

$$\tilde{y}_{(N-1)} \geq \eta^2. \tag{22}$$

The second condition requires that the correct model statistic y_ℓ must not be one of the top

k values, or equivalently that at least k of the remaining $N-1$ non-match models exceed the true model statistic. This condition is written as

$$\tilde{y}_{(N-k)} > y_\ell. \tag{23}$$

Applying these conditions we find:

$$\begin{aligned} P_{fd} &= \text{Prob}\{\tilde{y}_{(N-1)} \geq \eta^2 \text{ and } \tilde{y}_{(N-k)} > y_\ell\} \\ &= \int_{\eta^2}^{\infty} \text{Prob}\{\tilde{y}_{(N-k)} \geq y_\ell \mid \tilde{y}_{(N-1)} = y\}\, \tilde{p}_{(N-1)}(y)\, dy \\ &= \int_{\eta^2}^{\infty} \left(\int_0^y \text{Prob}\{y_\ell < u \,|\, \tilde{y}_{(N-k)} = u\}\, \tilde{p}_{(N-k)}(u)\, du \right) \tilde{p}_{(N-1)}(y)\, dy \\ &= \int_{\eta^2}^{\infty} \left(\int_0^y F_\ell(u)\, \tilde{p}_{(N-k)}(u)\, du \right) \tilde{p}_{(N-1)}(y)\, dy \end{aligned} \tag{24}$$

where in all cases the tilde indicates that $N-1$ non-match models are used and F_ℓ is the cumulative distribution function for the true model density expressed as in equation (18). The combined errors of false accept when the speaker is not known P_{fa} and falsely accepting the incorrect model from the known set of speaker models as above is referred to here as *false recognition*. The false recognition probability therefore equals

$$P_{FR} = \rho_1 P_{fd} + \rho_0 P_{fa}. \tag{25}$$

We can now observe performance for recognition by examining associated detection error tradeoff (DET) curves. We use the NIST parameter value $\rho_1 = 0.01$ for the false recognition probability (25) to produce the graphs in figures 13-14. In figure 13 we show missed detection and false recognition for the case where $\sigma = 1.5$ and the mean $\mu = 4.6$ set for single match Equal Error Rate performance equal to 2%. These curves show the effect on performance when k is changed from 1 to 3 to 6. As can be noted, there is no significant difference in performance as k varies, and in fact the curve for $k = 6$ effectively overlays the one for $k = 3$. This observation appears to hold for other values of N although not unexpectedly overall performance improves as N is reduced.

Figure 14 shows a comparison for $N = 200, k = 1$ and several means μ for fixed variance $\sigma^2 = 1.5$. The choices for $\mu = 3.5$ (red curve), 4.6 (green curve), and 5.2 (blue curve) correspond to EERs for single comparison detection of $6, 2$, and 1 per cent, respectively. Also indicated are the performance points along these curves arising from the NIST parameter settings ($\rho_1 = 0.01, \rho_0 = 1 - \rho_1, C_{md} = 0.91, C_{fa} = 0.09$) and ML equal cost and equal priors thresholds. As can be seen, these single-comparison test threshold settings are far from optimal for recognition purposes when trying to identify the correct speaker model from a larger set of speaker models. In fact, it is clear that these thresholds can be adjusted to greatly reduce the false recognition rate with minimal increase in missed detections. It suggests a dual strategy of using one threshold for the initial detection case

and then employing a different threshold for the N pairwise comparisons. Once again, the relative difference between these curves are not significantly altered for different values of N or k except that they move up or down.

6 Analyst Workload

We now consider the analyst workload for the recognition process when an analyst is required to judge the top-k matches from our set of speaker models. We assume the speaker is at first correctly judged not to be the claimed identity and the goal is to then correctly identify the correct speaker from our set of N speaker models or recognize that the speaker is not among our available speaker models. Based on the output statistics from the various models in the set, a strategy for the speech samples for those of the best k models which pass a threshold test to be passed to an analyst for further scrutiny. It is possible that not all of the top k (and in fact possibly none) will pass threshold. As shown above, the set size N and the number k that are to be considered both affect overall detection and false accept performance for the SID system. In particular, even though we've noted that detection performance can be improved with no degradation in false accept performance by increasing k, we examine here the impact on analyst workload for such a strategy.

Before presenting results, we emphasize that we are not considering the ability of a human analyst to select the correct speaker from the top k presented. Data on human performance for speaker identification is limited, and what little data there is has used naïve listeners, not professionals. Previous studies [1, 13] have shown that human performance was no better than machine performance provided the training and recognition samples came from the same source (e.g., landline, GSM cellphone, etc.). When the recognition samples came from sources different than those used in training, humans outperformed identification systems. This is believed to occur because the high-level features people use are more robust. Interestingly, anecdotal data [14] suggests that humans make different kinds of mistakes than computer-based SID systems, which suggests that the fusion of the two should be more effective than either alone.

Incorporating a model of human performance is beyond the scope of the current effort, so this paper instead focuses on any additional effort an analyst will need to devote to make a recognition decision based on different values of a number of variables, namely, the size N of the set of speaker models, top k selected for scrutiny, and single-match threshold settings. We decided to adopt (what we believe are) some reasonable assumptions regarding the workload process for an experienced analyst. The metric we utilize to measure workload is time spent by an analyst listening to recorded speech. Specifically, we measure workload as the number of hours of analyst involvement in hours per 100 trials, where a trial is defined as a single speaker being presented to the SID system for evaluation. In the simplest case ($k = 1$) we assume that the analyst spends approximately 30 seconds listening to the incoming speech and another 30 seconds listening to model speech. We surmise that if an analyst is required to listen to a speech segment to identify or verify an identify from $k > 1$

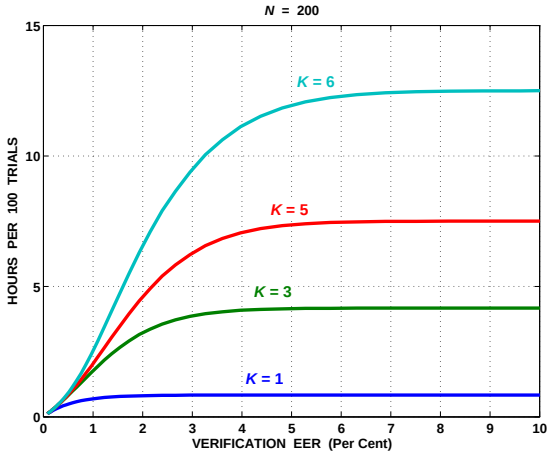

Figure 15: Workload (hours per 100 trials) based on set size $N = 200$ for detection strategies using $K = 1, 3, 5$, and 6.

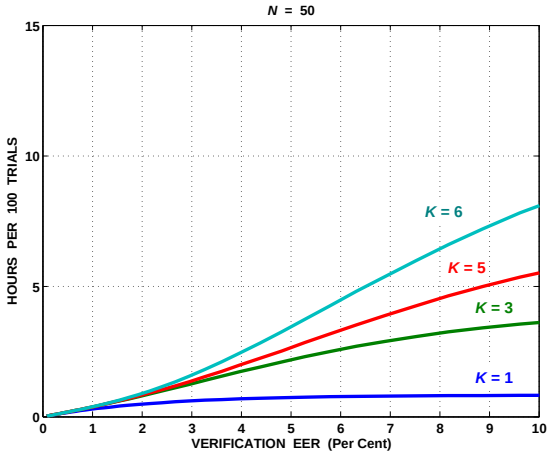

Figure 16: Workload (hours per 100 trials) based on set size $N = 50$ for detection strategies using $k = 1, 3, 5$, and 6.

stored speech segments, then he will listen to all k model comparisons before making a decision.

An assumption we make in the analysis is that the analyst uses a process of elimination in that he will listen through the entire list of k comparisons in order to first pare down the list in a first cut, eliminating a large portion but leaving a smaller number of speech samples with which to repeat the process. As k increases, we would expect that the analyst would need to repeat this paring process multiple times. For this study, we have assumed that if $k = 1$ or 2, a single stage suffices; if $k = 3, 4$ or 5 then a two-stage operation is needed. Our final assumption is that the second listening stage is more difficult than the first and will therefore require additional time. We therefore have assumed 1 minute for listening to each speech cut for the second stage. When $k > 5$ we assume a total of three passes allowing for a first cut to pare down to 3 and then a final cut comparing 2 while spending 1.5 minutes for this third cut.

Results presented in figures 15 and 16 not surprisingly show an increase in analyst workload as k increases. Comparing these two figures clearly indicates that increasing the size of the known speaker models for comparison can have a devastating effect on the analyst workload. A similar effect is seen as k is increased. The point of these results is not so much that the workload increases, which is expected, but how fast it increases. For example, where one might hope that using a strategy of passing the top $k > 1$ speaker matches might improve performance, it can very quickly overburden an analyst. Furthermore, since in some applications the size of known speaker models can grow, these curves indicate that in time the workload on the analyst will likely become untenable whenever $k > 1$. This is more readily illustrated in the plot in figure 17 where workload is plotted as a function of

k for $N = 50$ and 200 for several single-match EER rates of 0.5%, 1.0%, and 2.0%. (The jump in workload for $k = 6$ reflects the fact that a third analyst review is introduced at that level.) Taken together, this suggests that a methodology that can successfully segment the original set of known speaker models into smaller groups prior to performing detection can be very beneficial to speaker identification.

The problem, of course, is that if one divides a 200-person speaker model set into two 100-person speaker model sets using some speaker attribute, then that attribute must be highly accurate. Otherwise, when an unknown speaker is presented and the system segmentation attribute decides the speaker is in one group when he is actually in a different group, then the system is guaranteed to have made an error. The only speaker attribute known to have such high accuracy levels is gender, but even this attribute is not accurate 100% of the time. An examination of figure 17 shows the benefits of perfect segmentation into four groups of 50. For $k = 6$ and $N = 200$, at the 2% EER level it takes 6.5 hours to perform 100 trials. In comparison, the equivalent situation for $N = 50$ takes about 1 hour. The segmentation strategy appears inevitable for the large set sizes that could be expected in the future, but thus far this has not been a priority for the research community.

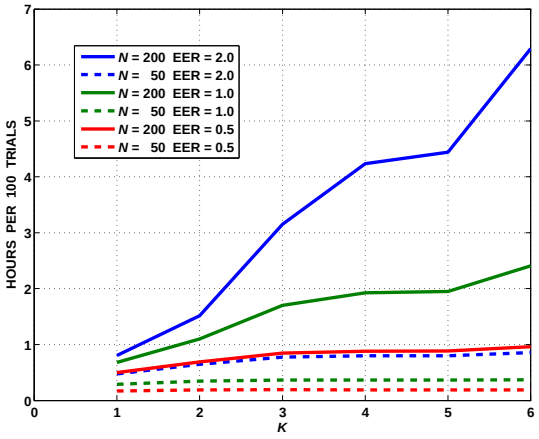

Figure 17: Workload (hours per 100 trials) as a function of k for several different set sizes and several values of EER ($\approx 2\%, 1\%$, and 0.5%).

Other options guiding the use of an analyst should be mentioned for completeness. One question is whether the analyst should always listen to all k cuts even if only a single one passes threshold. This reflects a certain lack of confidence in the SID system, and causes a predictable increase in analyst workload. Finally, we could instead use the SID system to simply rank order all N speakers and not threshold at all. The analyst then listens to k cuts every time. This is even more labor intensive, and is totally dependent on human judgments. If human labor is readily available or if the cost of a missed detection is enormous then either of these could be a viable option.

7 Summary and Conclusions

This article was motivated by the scarcity of analysis in the speaker recognition literature examining the implications of an application when a speaker is to be matched against a larger set of speaker models. Our discussion therefore addresses a very realistic situa-

tion: a speaker verification stage that is followed by matching against a speaker set when the verification fails. The mathematics of speaker detection is developed for maximum-likelihood, Bayesian risk criteria with either unequal priors or unequal costs and priors, and single-match (verification) equal error rates under the assumption that the statistical output of each speaker identification system is gaussian. This assumption is well justified since it appears consistent with real data collected and tested in the NIST SRE-08. We then showed the implications of large set sizes for current-day SID systems, particularly the large increase in the number of false accepts as the size of the set increases. This effectively renders these systems useless for real-world use. Analysts of such a system would spend virtually all of their time listening to false alarms. Even worse, it is generally recognized that when an automated system produces a very high error rate, analysts tend to ignore it. Our results suggest that the problem might be mitigated by using a dual threshold, one for the verification phase and a different threshold for the recognition phase. In effect, we accept a slightly higher missed detection rate in order to obtain a significantly lower false recognition rate. Figure 13 illustrates this tradeoff for a set size of 200 and an excellent detection error rate of 2%. For example, by changing thresholds (i.e., the point on the curve at which to operate) the analyst can work at the 1% false recognition point but with a cost of almost 30% missed detections. In the case that the a priori target probability is 1% this means that the analyst will listen to almost the same number of true hits (7) as false accepts (10) per 1000 trials.

We also considered the strategy of modifying the number of speaker cuts (represented by k) that an analyst should listen to for making a decision in the hope of improving overall system performance. By listening to the top-k cuts, we showed that detection rates do improve. Figure 10 showed that at the 2% verification EER point, raising k from 1 to 3 caused detection to go up to 96% from 92% thereby halving the missed detection rate. However, increasing k does not greatly reduce the overall error rate, which is dominated by false accepts when the a priori target probability is small ($< 1\%$). Furthermore, an even more significant problem with raising k is that based on our assumptions the overall analyst workload increases rapidly. This workload increase therefore renders this strategy as a niche technique: useful only in the rarest of situations such as when labor is readily available or the cost of a missed detection is extremely high.

Finally, we pointed out a possible approach for mitigating the performance degradation associated with large known speaker model set sizes, namely, segmenting the set in accordance with measurable speech characteristics to avoid being overwhelmed by false recognition issues. Several characteristics associated with speech patterns have been considered for this purpose (e.g., vocal tract length, dialect), but none except gender has yet proven accurate enough to make segmentation a practical option. This approach is necessary only because today the fundamental single-match verification technology can achieve at best an equal error rate of about 2% in formal evaluations. Thus both issues, namely, accurate segmentation and improving the core single-match verification technology, should be focus areas for SID development. For segmentation, a better understanding of the speech

characteristics that are most beneficial and reliable is needed.

With respect to core SID research using the same detection process (i.e., using the UBM) improvements are manifested by increasing the target distribution mean μ and decreasing the associated target distribution variance σ. A better understanding of what most influences these parameters should provide valuable insights for researchers. For example, the issue of why σ is greater than one is not yet well understood and needs to be investigated. As the core technology improves, greater speaker model set sizes can be supported without the use of segmentation. This of course will make feasible the practical use of even larger set sizes (again with additional segmentation). Inevitably, improvement in segmentation capabilities leads to the philosophical question of whether, when, or if the speech characteristics measurements used in segmentation should be integrated as part of the core SID technology.

Until some of the issues for improving SID capabilities are better addressed, the decision as to which strategy an analyst should adopt will remain difficult. As noted in this paper, analyst workload issues are significant. This study highlighted the key dimensions of this problem and the cost-benefit tradeoffs for some possible solutions.

References

[1] Alexander, A., *Forensic automatic speaker recognition using bayesian interpretation and statistical compensation for mismatched conditions*, Ph.D thesis (No. 3367), Swiss Federal Institute of Technology, Lausanne, Nov. 2005.

[2] Abramowitz, M. and I.A. Stegun, eds., *Handbook of Mathematical Functions*, Dover Publications, New York, 1970.

[3] Brummer, N., L. Burget, J. Cernocky, O. Glembek, F. Grezl, MK̇arafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, *Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006*, IEEE Transactions on Audio, Speech and Language Processing, Volume 15, Number 7, pp. 2072–2084, Sept. 2007.

[4] Campbell, J., *Speaker recognition: a tutorial*, Proceedings of the IEEE, Volume 85, Issue 9, pp. 1437–1462, Sept. 1997.

[5] David, H., *Order statistics*, Wiley and Sons, New York, 1981.

[6] Doddington, G.R., *A computer method of speaker verification*, Ph.D. dissertation, Department of Electrical Engineering, University of Wisconsin, 1970.

[7] Gish, H., M. Krasner, W. Russel, and J. Wolf, *Methods and experiments for text-independent speaker recognition over telephone channels*, Proceedings IEEE ICASSP, pp. 865–868, 1986.

[8] Gish, H., K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, and J. Wolf, *Investigation*

of text-independent speaker identification over telephone channels, Proceedings IEEE ICASSP, pp. 379–382, 1985.

[9] Goodman, F., analysis of NIST SRE-08 data, unpublished, 2008.

[10] Kay, S., *Fundamentals of statistical signal processing: detection theory*, Prentice Hall, New Jersey, 1998.

[11] Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker verification using adapted gaussian mixture models*, Digital Signal Processing, Volume lO, pp. 19-41, Jan. 2000.

[12] Rosenberg, A.E., *Evaluation of an automatic speaker verification system over telephone lines*, Bell System Technical Journal, Volume 55, Number 6, pp. 723–744, 1976.

[13] Schmidt-Nielsen, A., and T. Crystal, *Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data*, Digital Signal Processing, Volume 10, pp. 249–266, 2000.

[14] Van Leeuwen, D., private communication, 2008.

[15] Site: http://www.itl.nist.gov/iad/mig/tests/sre/index.html.

[16] Site: http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf.