Sequential Decisions for Faster and More Flexible Verification

Arun C. Surendran

Multimedia Communications Research Lab, Bell Labs, Lucent Technologies
600 Mountain Ave, Murray Hill NJ 07974.
email: acs@research.bell-labs.com

Abstract

Most speaker verification systems wait to collect a complete utterance from a speaker before making a decision. Faster verification can be achieved if decisions are made sequentially on smaller “chunks” of data. In this paper we present a sequential decision making algorithm in a connected digit application and discuss its properties. We show that sequential decisions, apart from requiring shorter utterances and fewer computations on the average, add another dimension of flexibility over current systems: within some limitations, they provide the ability to systematically tradeoff between the performance of the system and the amount of data needed to make a decision. Thus they make a speaker verification system work faster and be more flexible in real applications.

1. Introduction

Most speaker verification (SV) systems wait to collect an entire utterance and perform a single hypothesis test to arrive at a decision to accept or reject the claimed identity of the speaker. In many practical applications, especially in defense and intelligence, it is desirable to make a decision while the speaker is talking. In such scenarios, data is usually made available in a stream or small chunks. It would be beneficial to make sequential decisions on the smaller blocks of data as they become available rather than wait for the entire utterance. This is the issue we have addressed in this paper. We have devised a method to make decisions sequentially on speech segments at the digit level in a connected digit based SV system using hidden Markov models (HMM).

The focus of this approach is to make a decision with a desired level of confidence and performance as early as possible. Hence a sequential approach not only leads to a reduction in the the computing load of a system, but also leads to speed ups in their response times. Preliminary results indicate that we need about 7 digits per utterance to make decisions that are as reliable as using a fixed length of 10 digits - which gives a computational savings of about 30%. We also show how this approach brings the added ability to trade off between speed and performance thus providing flexibility and adaptability that is needed in practice.

2. Sequential Decisions

The idea in a sequential approach is to conduct a series of tests, each done when more data become available. At each step, a test is performed to find out if a decision can be made which meets some predefined degree of confidence; if it can, the result of the test is accepted; otherwise, the decision is postponed till the next step where more data become available. This process is repeated till a final choice is made. The goal is to design a test that can make confident decisions taking as few steps as possible on the average without sacrificing performance.

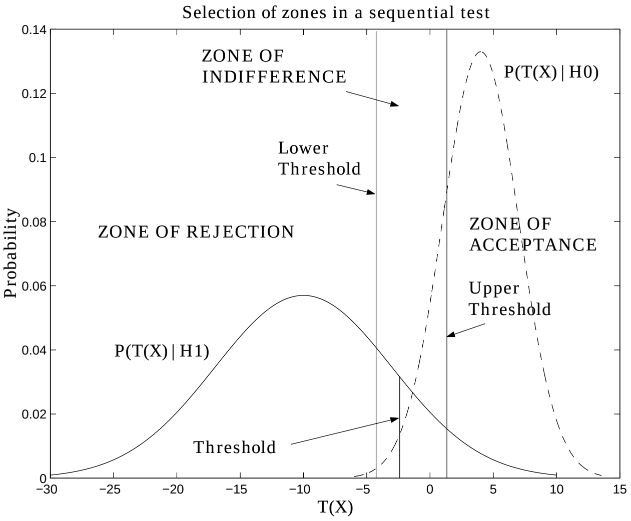

Figure 1: Three zones in a sequential test. Notice how the tail of both the distributions can lie outside the zone of indifference hence leading to early errors

In an SV system using a regular hypothesis test, given a spoken utterance $X = \{x_1, x_2, \ldots, x_N\}$ (where x_i can be composed of an individual sample or blocks of samples), a single test is performed that

$$\text{accepts claimed identity of speaker if} \quad T(X) > t \qquad (1)$$
$$\text{rejects} \quad \text{otherwise}, \qquad (2)$$

where $T(X)$ is the *test statistic*, and t is a threshold [1]. The choice of the threshold determines the performance of the system measured in terms of two errors: (1) false rejection (FR) of a speaker when the claimed identity is true (null hypothesis or H_0), and (2) false acceptance (FA) when it is not (alternate hypothesis or H_1).

In a sequential test [2], decisions are made in a series of steps, each time using the sequence of data points (or blocks) $X_m = \{x_1, x_2, ..., x_m, \},\ m < N$. At each step m, a test statistic $T_m(X_m)$, is computed and a test is performed that

$$\begin{aligned} \text{accepts speaker if} \quad & T_m(X_m) > B_m \\ \text{rejects speaker if} \quad & T_m(X_m) < A_m \\ \text{draw next sample} \quad & \text{otherwise}. \end{aligned}$$

where A_m and B_m are thresholds whose values represent the degree of confidence we have in the decision that we are making. $T(X) > A_m$ is called the *zone of preference for acceptance*, $T(X) < B_m$ is called the *zone of preference for rejections*, and $B_m \leq T(X) \leq A_m$ is called the *zone of indifference* [2]. The zone of indifference is usually defined such that X lies around the boundary between the two hypotheses. In this region, we are not confident about making a hard decision so we choose to postpone it till we get more data. We can see that having a wider zone of indifference may cause an increase in the average number of steps needed to make a decision (this number is called the *Average Sample Number (ASN)* [2]). The choices and consequences of these thresholds can be further understood from Figure 1. Depending on the overlap between the distributions of the null and alternate hypothesis, a narrower zone of indifference, while triggering quicker decisions, may lead to faultier choices earlier on in the test. This is because the decisions are made based on fewer samples, and hence may be more noisy and less reliable. From this we can hypothesize that, in general, tests that lead to smaller ASNs also lead to higher errors and vice versa. Thus the performance of sequential tests is measured not only in terms of FR and FA, but also in terms of ASNs. The aim of a sequential test is to achieve the desired level of performance using the fewest number of samples as possible.

A *sequential probability ratio test (SPRT)* is one where at step m, the ratio of the joint probabilities of the first m samples under the null and alternate hypotheses are used as test statistic:

$$T_m = \frac{P(x_1, x_2, ..., x_m, |H_0)}{P(x_1, x_2, ..., x_m, |H_1)}. \tag{3}$$

Such a test has many advantageous properties [2], the most important of which is that the thresholds A and B can be ideally selected solely based on the desired FA and FR of the overall test *without knowing the underlying distribution of the test statistic* [2]. Because of many factors (which we will discuss later) the choices of the thresholds do depend on the above mentioned distribution in most practical applications, and especially so in ours. Nevertheless, the algorithm we present here uses SPRTs.

3. Sequential Decisions for an SV System Based on Connected Digits

In speaker verification, the smallest sample - a single frame of data (usually based on 10ms of speech) - maybe too small to make a reliable decision. More reliable decisions can be made at a segment level, e.g. using phonemes (∼70ms) or words. Sequential tests can also be performed at an utterance level, where the speaker is prompted for a series of passwords or phrases till a decision can be reached. In this paper we make decisions at the word level. Even though there are scenarios other than connected digit based SV system where sequential tests may be more appropriate, we feel that this study is useful to demonstrates the principle and usefulness such tests.

Our sequential decision procedure is as follows: (1) When a block of data is received, it is analyzed to recognize the digit and identify the digit boundaries. (2) The digit is then scored using the corresponding digit models from both the claimed speaker and a speaker independent model set. (3) Based on the current score and scores of all the previously detected digits, a test statistic T_m is computed. (4) Based on the sequence of digits detected, and based on the degree of confidence desired, the upper and lower thresholds A_m and B_m are chosen. (5) If T_m falls below A_m, the claimed identity is rejected; if it is greater than B_m, the identity is accepted. Otherwise the system analyzes the next block of data. (6) Our experiments are done on a database where each user speaks a 10-digit string - so we have a maximum of ten samples (digits). If no decision can be made using the sequential test after the final sample, a hard decision of the type in Equation 2 is made to end the test.

In the following subsections, we will discuss the central issues of our algorithm: choice of the test statistic, the calculation of its distribution at each step in the test, and selection of the thresholds A_m and B_m.

3.1. Choosing the Test Statistic

We use the log-likelihood ratio (LLR) per frame for the partial utterance as the test statistic. This is the traditional metric used in SV systems ([1]). Since speech is a complex random process which is highly variable, and since our underlying models are assumed to be HMMs trained with a limited amount of data, the LLRs have a high degree of variability and are quite noisy. Hence using them does not give the optimal properties directly associated with an SPRT. Nevertheless, since it is not known how to design such an optimal test for speech signals, and since LLRs have been successfully used in SV systems [1], we continue to use them. However it is necessary to recast the metric in a form that can be computed sequentially based on individual digits, and whose distribution can be computed with reliability. To this end, the test statistic of an m digit string from Equation 3 can be written as

$$T_m = \frac{\sum_i^m s_i \hat{l}_i^t}{\sum_i^m \hat{l}_i^t}, \tag{4}$$

where s_i is the *average log-likelihood score per frame* of the word i and $\hat{l}_i^t$ is the length of the digit i in the test utterance. This information can be obtained from the decoder output.

3.2. Computing the Distribution of $T_m(X)$

As mentioned before, since our sequential test does not possess the optimal properties of an SPRT, knowledge of the distribution of the test statistic is needed to choose our thresholds. This can be done in a data-driven fashion given some ancillary data. There are many ways to calculate the distribution under the null and alternate hypotheses, each relying on the availability of either some speaker or imposter data or both [4]. One scenario that is realistic is to assume only the availability of data from other speakers even though they are not impostors (i.e. they do not speak the same pass phrase as the speaker) and use that data to compute the distribution of $T_m(X)$ only under the alternate hypothesis [3]. We adopt this procedure in this paper.

Since the underlying models are assumed to be HMMs, and since each s_i in Equation 4 can be written as the average of log-likelihoods of many frames of data (ignoring the state transition probabilities), using the central limit theorem, we can assume that they have normal distributions. The procedure to compute this distribution is not straightforward since the digit scores are correlated with each other. The parameters of this distribution are calculated from the mean and variances of each digit in the partial sequence and also the coefficient of correlation between each digit pair. The reader is referred to [3] for the complete procedure which is omitted here for lack of space.

3.3. Stopping Criterion - Selection of A_m and B_m

The procedure to select the lower and higher thresholds - A_m and B_m respectively - can be understood in conjunction with Figure 1. A_m and B_m can be interpreted as the degrees of confidence in our scores and they correspond to probabilities in our distribution of $T_m(X)$ under the null and alternate hypothesis. These values can each determine how early or late impostors and true speakers can be accepted or rejected, and they can be chosen based on the needs of the application. In our case, since we only know the distribution under the alternate hypothesis, our thresholds will depend purely on the confidence we have in rejecting a sample. For example, in our system A_m can be interpreted as the measure of our confidence in rejecting a true speaker. Since we do not know the distribution of T_m under the true hypothesis, we want to set a liberal lower threshold so as to not reject true speakers early. For our confidence in this choice to be higher than a particular level α, A_m has to be assigned a value such that $\int_{A_m}^{\infty} \mathcal{N}(t|\mu_m, \sigma_m^2)dt = \alpha$:

$$A_m = \mu_m + \sigma_m * \Phi^{-1}(\alpha), \tag{5}$$

where μ_m and σ_m^2 are the mean and variance of $P(T_m|H_1)$, and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-y^2/2}dy$. A good choice can be $A_m = 0.0$ which, while not rejecting almost 50% of the impostors till the last sample, does not reject a far higher percentage of true speakers.

4. Experiments and results

4.1. Database and System Description

The experimental evaluation is carried out using a database consisting of a digit string ("8 Z Z 8 2 2 3 5 Z 6") spoken by 45 speakers (20 male, 25 female), and recorded digitally over the telephone network using different handsets and under a variety of conditions. Each speaker provided 5 tokens of the string in a single training session, and up to 40 utterances over 20 testing sessions. For each speaker, 4 utterances of the same sex speakers are also used as imposter test data, which gives a maximum of 200 imposter utterances per speaker.

A 39 dimensional feature vector consisting of 13 cepstral coefficients and their $\Delta-$ and $\Delta\Delta-$ coefficients were derived per frame. The speaker and the background model both consists one HMM for each digit; the latter being built from the pooled data of all speakers, can be regarded as speaker-independent. Further details of the features and the models can be found in [3]. The average posterior equal error rate (EER) (where FA=FR) was 1.78% [3]. This is the best performance we can achieve given the test data or the exact distribution of the test statistic. Hence this performance can be never achieved in practice.

The digit strings were recognized and segmented using a good context dependent digit model. The boundaries thus obtained were used to pick the digits one by one in the sequential procedure. This procedure to obtain digit boundaries is not sequential, but since the state of the art digit recognition is extremely advanced and since digit boundaries can be obtained in practice with high degree reliability, this procedure is acceptable. In the future we will present results while performing recognition also in a sequential fashion. Such a sequential recognition procedure might give higher errors in a less restricted domain, e.g. 1000 word vocabulary system.

4.2. Results and Discussion

4.2.1. Evaluating our sequential algorithm

In this paper we have evaluated the performance of our system in four ways: (1) the variation of average error (FA+FR/2) with different choices of threshold, (2) the variation of ASN with different choices of threshold, (3) the variation of average error with ASN, and (4) the distribution of the number of digits needed to arrive at a final decision for a given pair (A, B).

These graphs tell us whether we can reduce the number of digits used while preserving a desired level of performance, and what kind of tradeoff we can achieve between performance and ASN. The results are shown in Figures 2 to 5.

4.2.2. ASN and Error vs. Thresholds

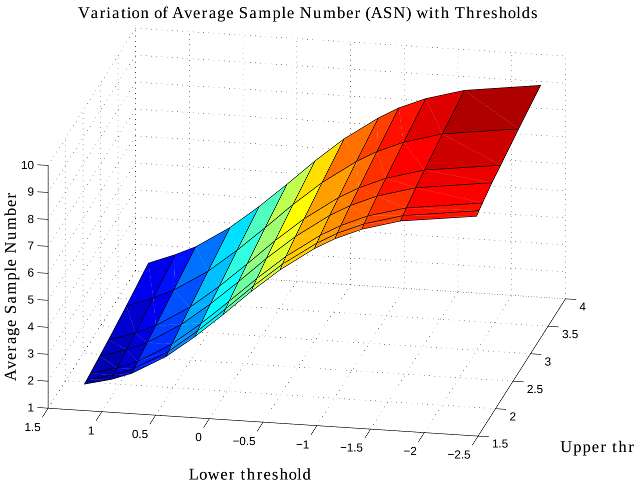

Figure 2: Average number of samples needed for decision for various choices of threshold. Values are averaged over 45 speakers.

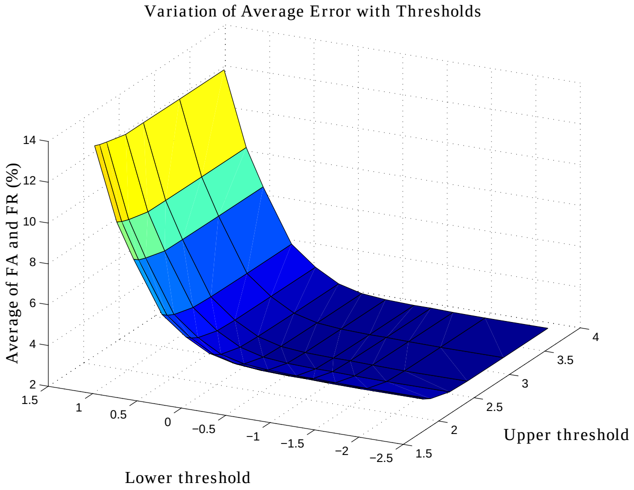

Figure 3: Average error - (FA+FR)/2 - for various choices of threshold. Values are averaged over 45 speakers.

Figures 2 shows how ASN varies with threshold. We can see that more liberal thresholds postpone the decision till later samples, and hence lead to larger ASNs. Since the distribution of T_m under the null hypothesis is not known, we have to adopt a liberal lower threshold. Hence A_m is set low and is varied over a wide range to study its effect. Figure 3 shows variation of average error with thresholds. As argued in Section 2, tighter thresholds can lead earlier decisions. Due to the high variability of speech signals, decisions made with limited data can trigger higher errors. This can be understood from Figure 1 where the tail of the distributions lie outside the zone of indifference leading to errors. There is a wide region in Figure 3 where the error is almost flat - this is a good region to situate our operating point. Given that we are operating at a given level of error, we look at the corresponding area in the ASN plot to choose a point where the the average number of samples is low. Lower ASNs lead to fewer computations and hence to quicker decisions. Thus, the two plots together highlight the flexibility of a sequential system - we have a systematic way to choose a tradeoff between average error and the amount of time taken to reach a decision.

4.2.3. *Average Error vs. Number of samples needed*

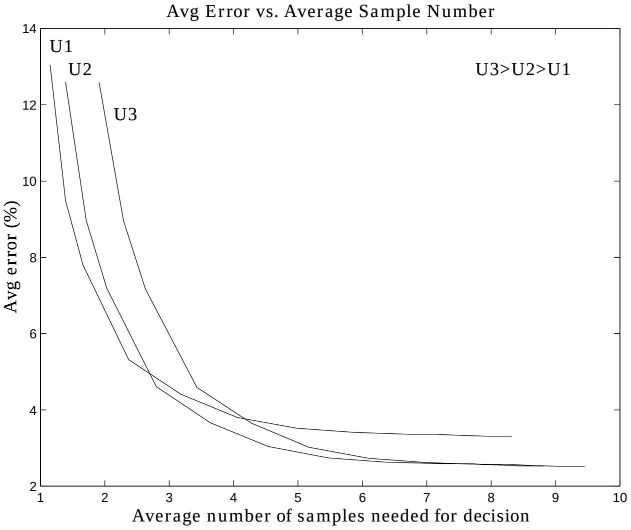

Figure 4: Average error vs. Average sample number. ASN is varied by varying the lower threshold. Each of the three plots shown are for different values of the upper threshold. Values are averaged over 45 speakers.

Since the goal of sequential tests is to reduce the number of samples needed for a decision without compromising performance, it would be appropriate to look at a plot of error versus ASN. This is shown in Figure 4. Each of the three lines are for a different higher threshold value ($B_m = U1, U2$ and $U3$). A_m is varied to vary ASN. The plots shows that errors are high for small ASNs, but they rapidly decrease till they reach a plateau. The height of the plateau signifies the best error that can be achieved for this upper threshold. As the upper threshold is increased the final error is lower, but it takes longer to achieve the lowest error. The lowest average error that is achieved is 2.52% (3.58% FA and 1.45% FR). The plots for U2 shows that an ASN of 7.2 digits gives an average error of 2.59%, and an ASN of 6.3 digits gives an error of 2.63%.

Figure 5 shows the distribution of the number of digits needed to reach a decision for a particular pair of thresholds. Notice that even though the mean is 4.39, a significant number of the decisions are made at a much lower level - with just one or two digits. This shows that even as little as one word is enough to make reasonable choices in speaker verification. The ASN for true speakers in this graph is 2.85 and for impostors it is 4.71. The higher ASN for impostors is due to the liberal lower threshold of $A_m = 0$ which assures that for almost 50% of impostors we will take all 10 samples to make a decision.

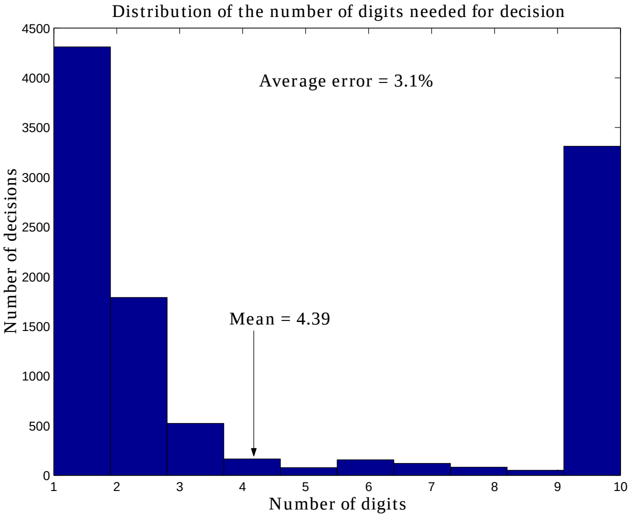

Figure 5: Distribution of the number of digits needed to obtain a decision when the lower and upper thresholds were (0,2.33). FA=3.57% and FR=2.64%, Avg error=3.09%. Results are for all 45 speakers.

In the future we propose to extend this work to text independent systems making decisions at the sub-word and frame level. Reliable online estimation of segments boundaries and computing the distribution of the test statistic at the frame and subword level are two of the many challenges that we have to address to successfully extend this approach.

5. Conclusion

In this paper we have introduced an algorithm which makes sequential decisions for speaker verification in a connected digit based system. We have demonstrated that instead of waiting to collect the entire utterance, testing the data sequentially can lead to, on the average, using fewer observations while producing an equally reliable test. We have shown that this not only leads to fewer computations and faster decisions, but also provides the flexibility to tradeoff between performance and speed. This technique is an important tool in creating faster, more flexible speaker verification solutions.

6. References

[1] C.-H. Lee, “A Tutorial on Speaker and Speech Verification”, Proc. NORSIG ’98, pp. 9-16, Vigso, Denmark, 1998.

[2] A. Wald, *Sequential Analysis*, John Wiley and Sons, Inc., 1947.

[3] A. C. Surendran and C.-H. Lee, “*A Priori* Threshold Selection for Fixed Vocabulary Speaker Verification Systems, Proc. ICSLP ’00, Beijing, China, 2000.

[4] J.B.Pierrot, *et. al.*, “A comparison of a priori threshold setting procedures for speaker verification in the CAVE project”, Proc. ICASSP, Seattle, 1998.