Proceedings of the Institute of Acoustics

THE ASSESSMENT OF SPEAKER VERIFICATION SYSTEMS

W Millar, J Oglesby, M Pawlewski & J G Tang

Speech and Language Technology, BT Laboratories, Martlesham Heath, IPSWICH.

1. INTRODUCTION

As interactive speech systems become more commonly accepted, there will be a growing demand for systems with security. Such systems are likely to incorporate some form of user authentication and it would be in keeping with the concept of an interactive speech service for these systems to employ speaker recognition by voice. Voice–based speaker recognition systems can be split into two categories, namely *verification* systems and *identification* systems.

In a speaker verification system, the user enrols by recording his voice in advance of using the system. The speech is analysed to extract features which are then used to construct models for use in a pattern classification system. At the same time the person is given a means of making an identity claim which is typically a personal identification number (PIN). When the user attempts to access the system, an identity claim is made by entering the PIN. New speech is requested and compared to that obtained at enrolment. A decision can then be made as to whether to accept or reject the identity claim based on the similarity of the new speech to the models.

In a speaker identification system a similar enrolment phase takes place but no PIN is needed. When an access attempt is made, the person will be required to speak to the system in order to be identified. The task is now much more difficult because the system has to identify the talker from all of those who have enroled. This requires a decision to select one from many (or to make a rejection) which is an inherently more difficult task.

As can be seen from the above descriptions, the operation of a speaker recognition system is inherently more complex than a speech recognition system. The majority of speech recognition systems are speaker independent and require no enrolment phase and hence no speech storage. The performance of a speaker recognition system must be consistent over time in order to provide the correct level of security. In contrast the performance of a speech recognition system can vary over time and the dialogue can be designed to compensate for this. Also more types of errors can occur and more parameters must be set during the design of a speaker recognition system. This in turn makes specifying the performance of a speaker recognition system more difficult. System developers will each have their own techniques for specifying performance (in order to monitor their own progress and track improvements), which means comparisons between different systems are difficult to make. Service providers face the same problems in comparing performance but also face problems at a more fundamental level in understanding the different parameters which may be specified and the relationship of these to their users' perceptions of a system.

It would be beneficial to system providers and algorithm developers alike for common performance measures to be agreed upon, which would allow systems to be easily specified and compared. The aim of this paper is to stimulate discussion in this area with a view to arriving at some commonly agreed metrics for future use. The paper begins by looking at the common measurements used on a speaker recognition system. For convenience we concentrate on speaker verification systems, since they are likely to be deployed in large numbers before speaker identification systems. Some of the performance measurements quoted by different manufacturers are then discussed. The paper then goes on to look at the factors which can affect the performance of verification systems and proposes that a standard corpus of speech be used. The dimensions of such a corpus are considered. Some remarks are made about the nature of the measurements which can be made from the corpus. Finally some conclusions and suggestions for adopting a standard mechanism in the assessment of verification systems are made.

2. COMMONLY USED MEASURES

A verification system can make two types of errors. When a valid user enters their correct PIN and is rejected by the system, this error is termed a true user rejection. The second type of error is when an imposter enters a valid user's PIN and is accepted by the system. This type of error is termed an imposter acceptance. In order to reduce the number of imposter acceptances, the system can be adjusted to require an extremely close match between the stored models and the speech spoken when an access attempt is made. This is done by modifying the threshold for accepting a user. As a result, the true users stand more chance of being rejected by the system if their speech does not compare as closely as required for some reason. Thus the two error measurements are not independent; reducing the number of imposter acceptances has the effect of increasing the number of true user rejections and vice versa. This is shown schematically in Figure 1.

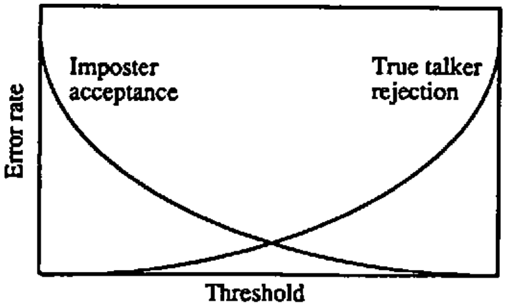

Figure 1. Verification errors.

The point where the percentage of true user rejections is the same as the percentage of imposter acceptances is termed the equal error rate (EER). Conceptually, in order to improve the performance of a verification system, system designers wish to do two things. They wish to increase the separation between the two curves so that the errors due to imposter acceptances and to true user rejections reduce and they wish to change the absolute value of the errors on the vertical axis. A consideration of the these two effects shows that the equal error rate is a good general purpose measure of system performance but there are a number of drawbacks.

The major drawback in using it is that the two error curves are drawn over varying values of thresholds. This requires access to the scores output from the classifier and thus must be undertaken by the algorithm developers or system manufacturers. It is unlikely that a service provider would have access to the algorithm to enable them to undertake the required analysis.

Another effect of using the EER is that the overall shape of the two error rate curves is ignored. The system may be required to operate away from the equal error point, eg a highly secure system could demand that no imposters were accepted. The EER would then give little information about the true user rejection figure at this level of imposter acceptance.

Some other measures which have been used to quantify performance are also based around the error curves. The false rejection rate for a given false acceptance rate can be found by finding the threshold value which gives the specified false acceptance rate and then finding the corresponding false rejection rate at this threshold. This gives an operating point away from the EER which is potentially more representative of a system. The EER rate for each talker can be calculated and then averaged. This tends to give a more optimistic estimate of performance since the system is tuned to each speaker.

From the above discussion it can be seen that although the EER is a generally accepted measure of a verification system, there are many factors which have to be taken into account when interpreting the value measured particularly when comparing systems. We shall consider more of these factors later; in the next section we look briefly at the diversity in performance claims.

3. MEASUREMENTS MADE BY MANUFACTURERS

To illustrate the diversity in measurements used by manufacturers a brief comparison will be made in this section. This has been done to highlight the variation rather than to recommend any one approach as being inherently better than the others.

Texas Instruments in their ICASSP paper [1] quote EER and show the graphs for a verification system in which the imposters know six of the other users' passwords. Voice Control Systems in their Eurospeech paper [2] quote error rates for imposter and true user rejections but it is not clear how these relate. The US Dept of Energy's Sandia Labs reported tests on a biometric verifiers in 1990 [3]. The results suggested that voice verification was less secure than other biometric approaches which led the manufacturers to suggest that the test procedures were incorrect! In particular, the requirement for users to utter the same password was cited as a problem. The results are reported for multiple (up to three) attempts to access the system. Bellcore report that the EER can double over a period of two weeks [4]. This highlights the problem of the time dimension in speaker verification systems in that the speaker is not static and one is interested in how the performance varies over time. This is of particular importance for true users. AT&T report results in terms of an average individual speaker EER [5]. It is assumed that all the speakers know the text of the test utterance. Voxtron report results for false rejections after three attempts, after exclusion of the first 10 days of use and after re-enrolment of users who had difficulty using the system after their first enrolment [6]. No false acceptance figures are quoted because none had been made.

From the above it can be seen that there is a wide diversity in the reporting of results, the conditions under which the measurements were made and even in the terminology used to define the errors. In the next section we consider in more detail the factors which affect the performance of verification systems.

4. FACTORS AFFECTING THE PERFORMANCE

In this section we discuss, in outline form, some of the parameters which can affect the performance of a speaker verification system. This is not an exhaustive list; its primary purpose is to highlight the number of factors which exist. Here we group the factors together under appropriate headings, with a view to considering mechanisms for quantifying their overall effects.

4.1 Environmental Effects

The factors which will be considered under this heading will be those which are due to the environment in which the speech data is recorded. In general these will be considered to be outside the control of the user and the system designer. However, it will be expected that a good system will be able to deal with and compensate for these effects. The type of effects which would fall into this category include the following:

1. *Background noise environment*. The background noise can range from the quiet environment that would be expected in a small room without any air conditioning to that in a busy street if someone was using a call box.
2. *Lombard effect*. The quality of the speaker's voice can change in response to the external environment.
3. *Microphone variations*. There will be a difference in quality between different transducers and between different positions of the same transducer. This is particularly true of the frequency response, but different constructions can also generate extraneous noise.
4. *Channel type and quality*. The channel could range from a high bandwidth low noise channel as may be used in a room access system to the narrow bandwidth high noise system more typical of that found in a mobile telephone network. This later channel type can also include additional signal processing components such as speech coders.

4.2 Data Capture Effects

At this level in the assessment of the algorithm we are moving away from the physical effects due to the environment in which the speech data has been obtained. The factors which need to be considered in order to ensure that "sufficient" data is used to provide accurate estimates of system performance are discussed. These factors will include:

1. *The number of talkers used to make the measurements*. The appropriate numbers of talkers will have to be used to ensure that the results are statistically significant. In addition, it is known that within the population as a whole there are two different categories of user, namely sheep and goats. The sheep will always manage to get a verification system working and the goats will not be able to do so. These groups need to be represented in an appropriate manner.

3. *Novice/expert users*. Different people will have different expectations and different attitudes to technology. This influence must be accounted for when assessing the perfomance of systems.

2. *Vocabulary.* The number of words used to perform the measurement has an effect on performance. In general it can be assumed that the more speech which can be made available, the better.

4.3 System effects

The parameters which will be identified in this section exist at a higher level in the system than those discussed in previous sections. These factors, which are under the control of the system designer and can be adjusted to modify the performance of the system, include:

1. *The nature of the utterance(s).* It can be conjectured that different lengths of utterances will perform differently in different verification systems, depending on the particular algorithm used. Additionally, other factors such as the phonetic structure and balance of the utterance, the familiarity of the utterance to the user, etc. will contribute to the overall performance.
2. *Prior knowledge.* In order to improve the performance of any verification system, users are given their own unique passwords. This creates problems in testing the imposter acceptances. How much information should the imposter be assumed to know? Some systems assume that the imposter does not know any of the unique passwords, other assume that the imposters may know a few of the passwords and yet others assume that the imposters know all the passwords.
3. *Number of repetitions.* The number of utterances collected during an enrolment phase and used to produce the reference models is known to have a significant effect on the performance of the system. The maximum number is constrained by human factors issues.
4. *Enrolment strategy.* This can be tailored to affect the performance significantly. For example, some manufacturers check utterances for confusability with those already enroled and remove these, others allow users to make trial calls before using the system.
5. *Decision strategy.* Many different decision strategies can be employed. Options which exist are to allow speech to be solicited as necessary until a decision can be made or alternatively, a fixed amount of speech can be obtained and a decision made on it.
6. *System performance over time.* Any verification system will be repeatedly used by customers over a period of time. Thus the performance of the system will need to be measured over time as well. This will allow the effects of updating strategies to be examined. This will also allow transient effects such as colds to be examined.

4.4 Implementation issues

As well as considering the performance of the algorithm in terms of accuracy, the implementation requirements need to be considered. Since verification systems encompass elements of speech recognition it will be assumed that if a system can perform speech recognition it will have the processing power to be able to perform speaker verification. A more important parameter for the implementation of the algorithm is the amount of storage required to hold the reference models which impinges on the storage space required and data transfer rates.

A further parameter which must be considered before implementation is the user acceptance of the enrolment strategy. This will be application specific.

4.5 Performance measurements

As we have seen in the sections 2 and 3, the type of measure which is used to determine the performance has an obvious effect on the perceived quality of the system, as well as the actual quality. As well as the problems with measuring and using the EER, care must be taken in specifying how the measurements are performed, since the performance can be seen to depend on the specified operating conditions.

5. TOWARD COMMON STANDARDS

Given the wide range of parameters which can influence the performance of a verification system and accepting that both algorithm developers and purchasers of systems will want to measure the performance, how then should we proceed toward a common method of assessing the performance? If we consider the problem it seems to split into two components: the measurements to be made and the data on which they are to be applied. We can now address each of these separately and make some proposals for future discussion.

5.1 A standard speech corpus

If we assume that we can derive some appropriate measurements, then speech utterances are required which are realistic of a verification task and which can be used to assess the performance in a meaningful way. A similar approach is used in the assessment of speech recognition systems particularly in the USA under the DARPA programme. The most immediate problems to be considered in designing this corpus are its size, its content and the recording conditions under which it is made. We shall consider each of these separately below.

5.1.1 Dimensions of the corpus. The three main considerations governing the size of the corpus are that it should be large enough to provide statistically significant results, it should reflect the composition of the user population and it should reflect typical patterns of use. The constraint on the size is imposed by the logistics of collecting, processing and storing a corpus of the size required.

The requirement of statistical significance can be addressed by undertaking error analysis using a normal approximation to the binomial distribution.

For a 95% confidence interval, $N = 3.84\,\frac{p(1-p)}{e^2}$,

where N is the number of test tokens, p is the error rate and e is the confidence interval.

e.g. if $p = 10.0\%$, $e = \pm 1.0\%$ then $N = \frac{3.84 \times 0.1 \times 0.9}{0.01 \times 0.01} = 3456.$

From this analysis it can be seen that 3500 test tokens are needed in order to ensure the true error rate is within 1.0 percentage points of the measured error rate 95% of the time if the true error is of the order of 10.0%. When comparing two algorithms, this confidence interval is a worst case value as in practice a tighter interval will result from using a better test for statistical significance such as McNemar's test [8] (for verification systems, this means fixing the threshold; we use the equal error point).

The requirement that the sample of talkers in the corpus reflect the proportion of sheep and goats in the general population can be addressed using a similar technique.

The third requirement of being representative of the task leads us to consider how the system will operate. A verification system providing security will be accessed more than once by users. For example a verification system used for telephone banking might be accessed at least once per month over a period of time as people make enquiries about their bank accounts. Similarly a voice controlled access system might be accessed at least once per day in order to allow access to a building. Thus true users are going to make multiple access attempts. This should be reflected in the speech corpus in that true users should have multiple utterances recorded over a period of time. We can arbitrarily say that there should be 35 repetitions of any individual utterance (corresponding to an access three times a month over a year). This means that a total of 100 true talkers are required in the database to give 3500 test tokens. In order to analyse the imposter rejection performance, if we allow an imposter to make only one access attempt, we need 3500 imposter recordings.

5.1.2 Content of the corpus. This is more difficult to specify and so we take guidance from the likely operation of a system and also how others have approached the problem. An identity claim is usually made orally through a PIN, thus it is appropriate to include all the digits, both in isolated form and as digit strings. If these are treated as passwords, then this provides a convenient mechanism for allowing all users to know everyone else's password. At the other extreme we need a password which is unique to each individual. This can be used to assess the additional security due to each user having a unique password although it provides no indication of whether the password remains secure in use. Convenient words to use for this task can be surnames or street names. Since some manufacturers allow each person in the test population to know the password of a subset of the population, it is appropriate to include in the vocabulary a small number of words which are known to a small group of people, say one or two words shared between five people.

This results in an initial vocabulary size of 10 digits plus say five digit strings of different length plus a surname plus say two other words. This is a total of 18 vocabulary items per recording session which is large enough to be meaningful yet small enough to be manageable to collect in each session.

5.1.3 Recording conditions. This is almost an emotive subject when recording talkers. There is much debate over the merits of recording in a controlled environment such as a silence cabinet or a good quality recording studio as against recordings made in office or home environments. When a system is deployed, it is unlikely that it will be possible to control the operating environment and so it is realistic to make the recordings in the same environment as that in which the system will be used. If it is deemed to be desirable to have some control over the background environment, it may be possible to make use of techniques for simulating background noise levels as have been developed and agreed by those involved in measuring the performance of speech transmission systems for telephony applications.

5.1.4 Additional factors. It may be desirable to include in the corpus groups of people which humans have difficulty separating, eg familial groups, uncommon accents, etc. While it is not clear how confusable machines might find these people, inclusion of them in the corpus is likely to give some indication of the discriminating capabilities of the algorithms in the systems.

5.2 Standard measurements
Devising and agreeing some standard form of measurement is seen as a more difficult task.

Developments of the imposter acceptance and true user rejection curves to encapsulate much more information than is in the EER is one possibility. One could consider quoting the EER together with points on the curves at some proportion of the EER. This would give other points on the curves which will help an evaluator assess the performance by providing some information about the shape of the curves.

An alternative or possibly complementary mechanism is to adapt the concept of a capability profile for speech recognition [7] to include verification specific parameters. These might be response time, influence of password knowledge, whether the algorithm requires few or many utterances for verification, whether it works on high quality speech etc. This would allow the inclusion of parameters which are important to system developers and service providers.

6. CONCLUSION

This paper has considered some but not all of the factors which can affect the performance of a speaker verification system. Because there is such a wide range of factors which can influence the performance, it is difficult to compare systems. This creates problems for both algorithm developers and service providers alike. In order to address this problem, the paper has proposed using a standard corpus of speech and standard measurements which can be made on the corpus. The speech corpus has been designed to encapsulate many factors which can influence performance and the measurements have been proposed to encompass both algorithmic and system issues.

The paper has been written with the aim of stimulating discussion in this area, since this is seen as important topic for consideration by those with in an interest in providing speaker verification systems.

7. REFERENCES

[1] J M NAIK, L P NETSCH & G R DODDINGTON, ' Speaker verification over long distance telephone lines ', ICASSP 89 Glasgow p524-527.
[2] A K HUNT, ' New commercial applications of telephone network based speech recognition and speaker verification' , Eurospeech 91.
[3] PIN news vol 6 no 6 July 1990.
[4] PIN news vol 7 no 3 March 1991.
[5] A E ROSENBERG, C-H LEE, S GOKCEN, ' Connected word talker verification using whole word hidden Markov models ', ICASSP 91 Toronto p381-384.
[6] B BIGELOW, ' Speaker verification over the telephone - a technical perspective of practical applications ', Speech Tech 88 p119-122.
[7] R K MOORE, ' Recognition, the stochastic modelling approach ', Chapter 7, Speech Processing, Magraw Hill 1990.
[8] L GILLICK, S J COX, ' Some statistical issues in the comparison of speech recognition algorithms ', ICASSP 89 Glasgow p532-535.