Performance Evaluation in Open-Set Speaker Identification

A. Malegaonkar[1*], A. Ariyaeeinia[2]

[1]Auraya Systems Pty. Ltd. Sydney, Australia
[2]University of Hertfordshire, College Lane, Hatfield, Hertfordshire, UK
amit.malegaonkar@auraya.net, a.m.ariyaeeinia@herts.ac.uk

Abstract. The concern in this study is the approach to evaluating the performance of the open-set speaker identification process. In essence, such a process involves first identifying the speaker model in the database that best matches the given test utterance, and then determining if the test utterance has actually been produced by the speaker associated with the best-matched model. Whilst, conventionally, the performance of each of these two sub-processes is evaluated independently, it is argued that the use of a measure of performance for the complete process can provide a more useful basis for comparing the effectiveness of different systems. Based on this argument, an approach to assessing the performance of open-set speaker identification is considered in this paper, which is in principle similar to the method used for computing the diarisation error rate. The paper details the above approach for assessing the performance of open-set speaker identification and presents an analysis of its characteristics.

1 Introduction

In general, speaker identification is defined as the process of determining the correct speaker of a given test utterance from a population of registered speakers [1-2]. If this process includes the option of declaring that the test utterance does not belong to any of the registered speakers, then it is specifically referred to as open-set speaker identification. An inherent feature of this process is that it provides the possibility of establishing individuals' identities without the need for any identity claims. This in turn offers the capability for enhancing the security aspect of speaker verification through the screening process. Such screening may be required at the enrolment phase to minimise the possibility of multiple identity acquisition, or deployed at the verification stage to increase the capability to detect access attempts by impostors.

---

[*] During the course of this work, Malegaonkar was with the University of Hertfordshire.

Given a set of registered speakers and a sample test utterance, this task is defined as a twofold problem [3]. Firstly, it is required to identify the speaker model in the registered set that best matches the given test utterance. This is the process of identification. Next, it is required to determine if the test utterance is actually produced by the best matched speaker or it is originated by a speaker from outside the registered set. This is the process of verification. When the speaker is not required to provide an utterance of a specific text, the task is called Open-Set, Text-Independent Speaker Identification (OSTI-SI). In the literature, it is acknowledged that OSTI-SI is the most challenging class of speaker recognition [3-4]. A factor influencing the complexity of OSTI-SI is the size of the population of registered speakers. In theory, as this population grows, the confusion in discriminating amongst the registered speakers is likely to increase and therefore the number of incorrect identifications is likely to increase as well. The growth in the said population also increases the difficulty in confidently declaring a test utterance as not belonging to any of the registered speakers, when this is indeed the case. The reason is that, as the population size grows, the possibility of a voice originating from an unknown speaker being very close to one of the registered speaker models increases. The problem of OSTI-SI is further complicated by undesired variation in speech characteristics due to anomalous events. These anomalies can have different forms ranging from the communication channel and environmental noise to uncharacteristic sounds generated by the speakers. The resultant variation in speech causes a mismatch between the corresponding test and pre-stored voice patterns. This can in turn lead to degradation of the OSTI-SI performance.

Conventionally, the evaluation of OSTI-SI performance has been based on separate representations of the identification and verification effectiveness. However, for the purpose of comparing the performance of different systems, it is thought to be beneficial to consider a measure of performance for the complete process.

2 Evaluation Methodology

Figure 1 summarises the process of open-set, text-independent speaker identification (OSTI-SI). As shown in this figure, the given test utterance is assigned to the speaker model that yields the maximum similarity over all speaker models in the system, if this maximum likelihood score itself is greater than the threshold. Otherwise, it is declared as originated from a non-registered speaker. It is evident from the above description and Figure 1 that three types of error are possible in this process. These, which collectively define the conventional approach to evaluating the performance of OSTI-SI, are described as follows.

- A test utterance from a specific registered speaker, showing its highest similarity to the reference model for another registered speaker.
- Assigning the test utterance to one of the speaker models in the registered set when it does not belong to any of them.
- Declaring the test utterance, which belongs to one of the registered speakers, as originated from a non-registered speaker.

For the purpose of this paper, these types of error are referred to as OSIE, OSI-FA and OSI-FR respectively (where OSI, E, FA, and FR stand for open-set identification, error, false acceptance, and false rejection respectively).

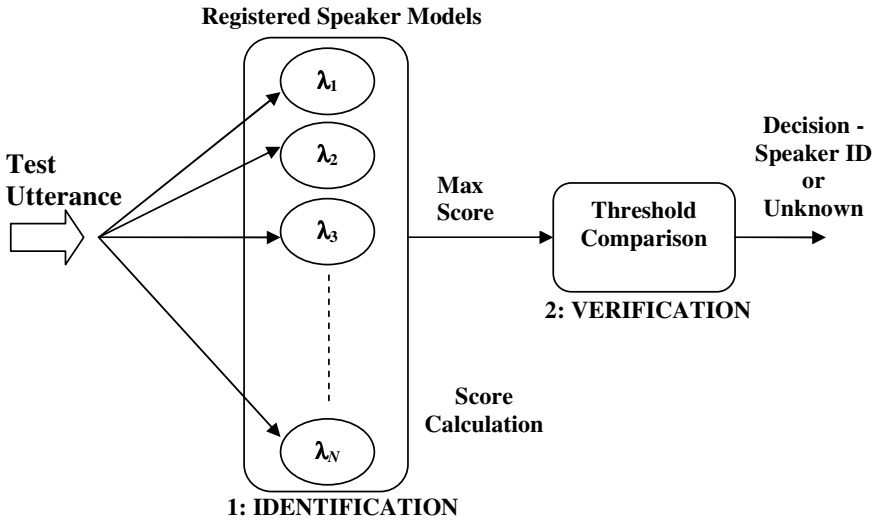

Fig.1. Overview of the open-set, text-independent speaker identification process

It is clear that the identification process is responsible for generating OSIE whereas, both OSI-FA and OSI-FR are the consequences of the decisions made in the verification process. It should be noted that an OSIE in the first stage would always lead to an error regardless of the decision in the second stage. Therefore, in evaluating the performance in the verification stage, it is important to discard the false speaker nominations received from the first stage (when the actual speakers are within the registered set).

As indicated earlier, an alternative approach to evaluating OSTI-SI is that based on observing the complete performance of the system. For this purpose, the operations involved in OSTI-SI are considered hidden in a box as shown in Figure 2. The system input is a test utterance and the output can either be a decision giving the identity of a speaker or a decision declaring that the test utterance does not belong to any of the registered speakers (shown as Unknown).

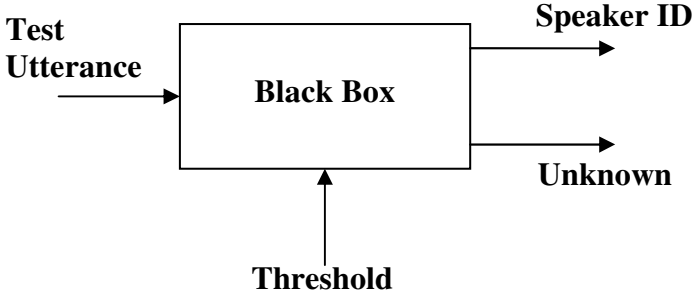

Fig. 2. Proposed basis for the evaluation of OSTI-SI

With such a configuration, three types of error can be recorded for a given threshold as follows.

- A test utterance from a registered speaker is associated with an incorrect speaker identity.
- A test utterance from a registered speaker is declared to have been produced by an unknown speaker.
- A test utterance from an unknown speaker is associated with a registered speaker identity.

In this study, the above errors are referred to as Mislabelling (ML), False Rejection (FR) and False Acceptance (FA) respectively.

In order to obtain the overall performance of OSTI-SI, a measure for combining all the possible types of errors is required. Motivated by the method used for calculating the diarisation error rate [5], an appropriate measure that can be proposed for this purpose is that of Accumulative Error Rate (AER). This is expressed as

$$\mathrm{AER}(\varsigma) = 100 \times \frac{\mathrm{ML}(\varsigma) + \mathrm{FR}(\varsigma) + \mathrm{FA}(\varsigma)}{T}, \tag{1}$$

where ς is the adopted threshold, T is the total number of tests, and X(ς) is the number of decision errors of type X for the adopted threshold ς. It should be noted that all three error types identified in this methodology, and hence AER are dependent on the decision threshold. Therefore, if required, equation (1) provides a means for setting the threshold such that the total error in OSTI-SI is minimised.

3 Experimental Investigations

This section details the experimental work conducted in order to further analyse the characteristics of the proposed evaluation methodology for OSTI-SI.

3.1 Speech Data

The speech data adopted for this investigation is based on the dataset used for the 1-speaker detection task of NIST SRE 2003 database. The protocol used in this work is based on that devised in [3]. The overall configuration of this dataset is given in Table 1.

3.2 Speech Features and Speaker Representation

Each speech frame of 20ms duration is subjected to pre-emphasis and then analysed to extract a 12^{th} order linear predictive coding-derived cepstral (LPCC) feature vector at a rate of 10ms. The static features are mean normalised. The first derivative parameters are also adopted and are based on the polynomial fit over 15 frames. These parameters are appended to the static features.

In this work, each registered speaker is represented by an adapted Gaussian Mixture Model (GMM) with 1024 components. For this purpose, a gender independent universal background model (UBM) is first obtained by pooling two gender dependant UBMs. The models for the registered speakers are then obtained using a single step adaptation of the gender-independent universal background model [6-7].

Table 1. Configuration of the dataset

	Female	**Male**
Registered Speakers	80	62
Registered Tests	767	526
Non-registered Speakers	93	48
Non-registered Tests	893	515
Speakers for Universal Background Model (UBM)	58	42
UBM Data Length	4.8 hrs	3.3 hrs

3.3 Results and Discussions

The results of this study in terms of ML, FR, FA, and AER as a function of the threshold are presented in Figure 3. In this figure, MLR, FAR and FRR are the rates of ML, FA and FR errors respectively. As observed in this figure, ML and FA errors decrease by increasing the threshold whereas FR error shows an increasing trend with an increase in the threshold. Variation in AER shows an interesting trend. This curve shows a distinct point of minima which is referred to as the point of Minimum-AER (M-AER). This point represents minimal total incorrect decisions in OSTI-SI. Hence this point can be an appropriate basis for setting the system threshold for OSTI-SI. Moreover, this measure is useful in comparing the performance of alternative OSTI-SI systems. It can also be observed that the largest component of errors at M-AER point is FR and the increase in FR is associated with reduction in ML decisions.

As discussed earlier, the individual processes of identification and verification in OSTI-SI are responsible for generating the overall decision errors in OSTI-SI. In addition to observing the overall performance of these processes, the analysis of the individual processes is certainly useful for understanding the limitations of the techniques used in implementing these processes. This is further useful for developing suitable techniques in order to improve the performance of either of the two specific processes, and hence OSTI-SI.

The variation in OSIE, OSI-FA and OSI-FR with the threshold is shown in Figure 4. The results in this figure are based on the same speech material as that used for the plots in Figure 3. It is observed that, in this analysis method, OSIE is independent of the threshold. The performance of the verification stage is then evaluated at the point of equal OSI-FA and OSI-FR. This is the point of Equal Error Rate (EER) for the verification stage and is referred to as OSI-EER.

Comparing figures 3 and 4, it is observed that FAR and OSI-FAR curves are exactly the same. The reason for this is that the tests originating from non-registered speakers are handled in a similar manner, regardless of whether the internal processes are considered independently or jointly.

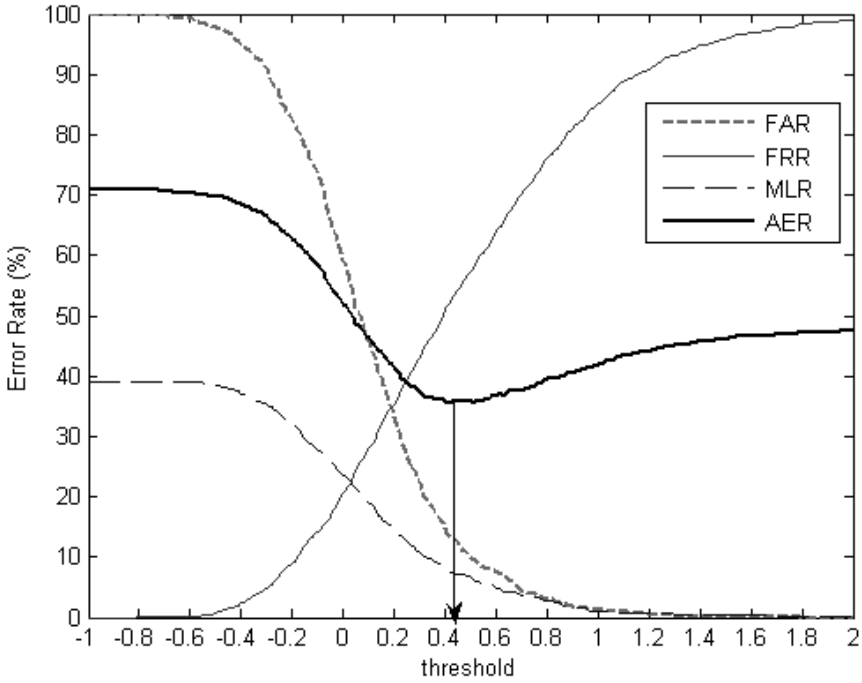

Fig. 3. Variation of error rates in OSTI-SI with the threshold

It can also be noted that OSI-FRR curve is different from FRR. The reason for this difference is that (as indicated earlier) in evaluating OSI-FRR, the tests resulting in OSIE are discarded. It is also observed that MLR curve has a characteristic similar to FAR curve. The reason for this is that, like FA decisions, ML decisions are generated due to acceptance decisions in the verification stage. Lastly, it should be noted that M-AER point is different from OSI-EER point and these are associated with different thresholds.

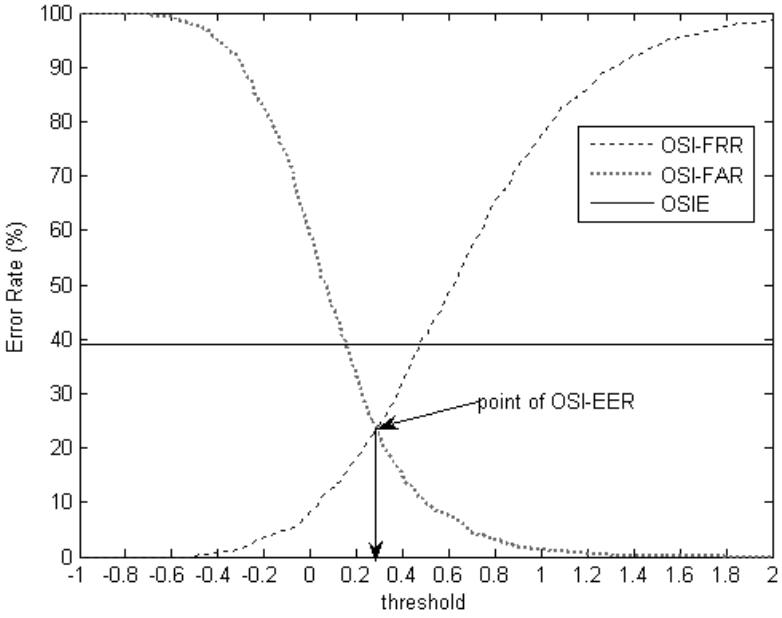

Fig. 4. Variation in OSIE, OSI-FA and OSI-FR with the threshold.

4 Conclusion

An alternative methodology for evaluating the performance of open-set, text-independent speaker identification (OSTI-SI) has been investigated. The introduction of this methodology is motivated by the approach commonly used in computing DER (diarisation error rate). It involves a holistic approach to the analysis of the performance in OSTI-SI rather than the independent consideration of the effectiveness in each of the two stages of the process (i.e. identification and verification). For this purpose, the use of three measures of the overall performance in OSTI-SI, i.e. mislabelling (ML), false acceptance (FA) and false rejection (FR) are considered. The integration of these measures has been achieved through the introduction of a metric termed Minimum-Accumulative Error Rate (M-AER). It has been shown that ML, FA and FR are all influenced by the threshold level adopted in open-set identification, and that it may not be possible to achieve equal rates of these errors using a single threshold level. However, it has been demonstrated that the threshold can be set such as to minimise the Accumulative Error Rate. The Minimum-Accumulative Error Rate provides a valuable basis for comparing the overall effectiveness of different open-set speaker identification systems. It has also been argued that, along with such a combined evaluation approach, the independent analysis of the individual processes involved in OSTI-SI can also be beneficial.

5 References

1. Pillay, S., Ariyaeeinia, A., Sivakumaran, P., Pawlewski, M.: Open-Set Speaker Identification under Mismatch Conditions. Proc. 10th Annual Conference of the International Speech Communication Association (Interspeech'09), 2347-2350 (2009).
2. Ariyaeeinia, A., Fortuna, J., Sivakumaran, P., Malegaonkar, A.: Verification Effectiveness in Open-Set Speaker Identification. IEE Proceedings Vision, Image and Signal Processing, 153, Issue 5, 618-624 (2006).
3. Fortuna, J., et. al.: Relative effectiveness of score normalisation methods in open-set speaker identification. Proc. the Speaker and Language Recognition Workshop (Odyssey), 369-376 (2004).
4. Singer, E, Reynolds, D.: Analysis of multi-target detection for speaker and language recognition. Proc. the Speaker and Language Recognition Workshop (Odyssey), 301-308 (2004).
5. Anguera Miró, X.: http://www.xavieranguera.com/phdthesis/node108.html. PhD Thesis, Speech Processing Group, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, (2006).
6. Reynolds, D., et. al.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10, no. 1-3, 19-41 (2000).
7. Fortuna, J., et. al.: Open set speaker identification using adapted Gaussian mixture models. Proc. Interspeech, 1997-2000 (2005).