BEYOND DODDINGTON MENAGERIE, A FIRST STEP TOWARDS

Juliette Kahn\textsuperscript{1,2}, Solange Rossato\textsuperscript{2}, Jean-Francois Bonastre\textsuperscript{1}

\textsuperscript{1}University of Avignon, Laboratoire Informatique d’Avignon (LIA) , France
\textsuperscript{2}University of Grenoble, Laboratoire Informatique de Grenoble (LIG) , France
\{juliette.kahn\}, \{jean-francois.bonastre\}@univ-avignon.fr, solange.rossato@imag.fr

ABSTRACT
During the last decade, speaker verification systems have shown significant progress and have reached a level of performance and accuracy that support their utilization in practical applications, including the forensic ones. This context emphasizes the importance of a deeper analysis of the system’s performance over basic error rate. In this paper, the influence of the speaker (his/her ‘voice’) on the performance is studied and the effect of the model (the training excerpt) is investigated. The experimental setup is based on an open source system and the experimental context of NIST-SRE 2008. The results confirm that the lower performances are obtained from a reduced number of speakers. Even more than speaker factor, speaker verification system performances are shown to be highly dependent on the voice samples used to train speaker models.

Index Terms— speaker verification, performance evaluation

1. INTRODUCTION
Over the last decade, automatic speaker verification systems (SVS) have been assessed regularly by the National Institute of Standards and Technology (NIST) [1]. The evaluation focuses on text-independent speaker detection and proposes a common experimental protocol and a stable set of evaluation rules. Even if the task difficulty is changing between the years, the NIST campaigns clearly show a drastic progress in terms of performance during the last years. The level of performance reached by the systems has become suitable for a large set of practical, commercial applications. Many applications are already available or planned for the next future, including the forensic ones. This context underlines the importance of a deep analysis of the system’s performance while the performance is currently usually assessed only through basic error rate.

The performance of such a system is measured using two kinds of errors. A false acceptance (FA) happens when an imposter is accepted by the system. A false rejection (FR) consists of rejecting a valid identity. Both error rates depend on a threshold used in the decision making process. Among the measures used to compare system performances, detection error trade-off (DET) curve [2], Equal Error Rate (EER) and Decision Cost Function (DCF) are usually used. The DET curve is obtained by plotting on a normal deviate curve the FA rate as a function of the FR rate. The EER corresponds to the operating point where FA ratio = FR ratio when the DCF corresponds to a specific operating point, described by the weight tied to each error (FA and FR) and the prior probability of these errors. These measures are computed on a large amount of data from many speakers. For the 2008 NIST evaluation, systems reached encouraging results (about 1.04% of EER, for long training in [3]). Though EER is an efficient measure to compare systems, further analyses are needed to better understand the topology of the errors. Doddington et al [4] studied the behavior of the speakers with respect to 12 automatic verification systems. They distinguished 4 types of speakers, illustrated by a ‘menagerie’. Sheeps correspond to the default speaker type (low FA, low FR). Goats are speakers who generate a disproportionate false rejection rate. Lambs correspond to speakers who generate a disproportionate false alarm rate. Finally, Wolves correspond to speakers that are likely to be mistaken for an other speaker. Doddington et al demonstrated that the topology of the errors depends on speakers. In addition to the speaker specificities, the voice samples used for training and testing could play an important role. Speaker information may depend on the number of speech frames selected by the system, even if the global duration of the file is appropriate, or on the phonetic content: some speech segments may be more relevant for speaker verification than others, even for a similar number of selected speech frames.

This paper focuses on the role of training signals according to the speaker and the choice of the voice sample. We present in section 2 the database and the system before analyzing, in section 3, the FA and FR rates according to the speakers. Section 4 proposes an analysis of the errors depending of the training excerpt in order to estimate the effect of the voice sample on the performance. Finally, we propose some discussion and ideas for future work in section 5.
2. DATABASE AND SYSTEM

The corpora is part of the male 2008 NIST-SRE telephone speech database. Most of the data are in English, but some conversations were recorded in other languages. The segment duration is approximately 2.5 minutes (condition short2-short3 in NIST protocol). This original condition of the NIST-SRE08 protocol is referred as NIST08 in this paper. NIST08 contains 221 different speakers, 11,636 non-target trials and 874 target trials.

The speaker verification system is the open source toolkit ALIZE/SpkDet [5]. This system is regularly assessed during the NIST speaker recognition evaluation. It is based on the UBM/GMM approach and it includes a latent factor analysis inter-session variability modeling [6]. For simplicity reasons, no score normalization is applied as the score normalisations show little effect on the performances (figure 1).

3. PER SPEAKER ERRORS

We examine the performance depending on the target speaker. Compared to Doddington et al [4], only the Lambs and the Goats are studied as we focused of the training.

3.1. Protocol

In order to maximize the number of target trials for each speaker, a leave-one-out scheme was implemented. For each speaker, a speaker model is trained using a speech sample while the other available samples of this speaker are used as target tests. This process is repeated for each speech segment available. This protocol is referred in this paper as M-08. 50 speakers of NIST08 pronounced less than two speech segments and are removed. Therefore M-08 includes 171 speakers with 816 excerpts, which means between 3 to 15 voice samples per speaker. Each model computed from a given training excerpt is compared with 801 to 813 non-target tests and 2 to 14 target tests. It results in a total of 661,416 non-target and 3,624 target trials. We first analyze the scores distribution for non-target trials on the one hand and for target trials on the other hand. Next, the error rates per speaker (FA and FR respectively) are sorted and cumulative curves are presented.

3.2. Results

3.2.1. Global distributions of scores

With M-08, the EER is 8.9%. This result is slightly higher than the one obtained in [3] (about 6%) may be due to the corpora and the lack of score normalization. Figure 2 shows the non-target and target score distributions with a Gaussian estimate of which distribution for comparison.

![Fig. 1. DET Curves without normalization and with ZT, Z and T normalizations (NIST08, male, english only)](image)

![Fig. 2. Non-target scores distribution ($\bar{x}=0.2559; \sigma=0.0922$) (top) and Target scores distribution ($\bar{x}=0.0091; \sigma=0.1720$) (bottom)](image)

3.2.2. Lambs

The global FA rate is 8.9%. When calculating FA rates per speakers, a range between 0.15% (speaker 71) and 29.58%
(speaker 200) is observed. In order to analyze more deeply this variation, the FA rates per speaker are sorted in ascending order. The cumulative distribution is shown in figure 3. The errors are far from being equally distributed across speakers. 50% of the errors are related to only 45 target speakers (26% of the 171 target speakers), and the ten worst speakers (i.e. 6% of the speakers) account 17% of the errors. Such a tendency corresponds to speakers labeled as lambs in Doddington et al.

3.2.4. Discussion

If a large range of FA and FR rates is observed, it is worth noting that there is no real boundary between the lambs and the other speakers when the goats seem to be more separated. Nevertheless, this result may be explained by the larger number of non-target trials compared to the target ones (661,416 non-target trials vs 3,624 target trials). Except for one speaker (speaker 112), the target speakers with the highest FA rates are not the same than those with the highest FR rates. Concerning the speaker 112, 3 speech segments contain mainly non-speech frames and grunts while the recording quality of the 2 other excerpts is poor. This observation points out that the speech excerpt itself may also play an important role.

4. CHOICE OF THE TRAINING EXCERPT

The speaker behavior pattern is studied regardless of the excerpts used thus far. The speech excerpt itself may have an important impact on the system performance, shadowing some other factors. In order to evaluate this point, we analyze the performance variation considering the training excerpt selected for a given speaker model.

4.1. Protocol

In this experience, we select the best and the worse training files per speaker, among all the speech excerpts available for this speaker. The performances obtained with both training excerpts are compared to the one obtained using the training file defined in NIST08 used as baseline. The performance is estimated globally (like in NIST protocol), because of the database size limitations.

To ensure a fair comparison, each model issued from one of the training excerpt is compared to the same test trials list. Also, the speakers with only two speech excerpt are discarded. Finally, the training excerpts were not used as test trials (which could occur when selecting each available voice excerpt as a potential training file).

These constraints give an experimental protocol with 511 target trials and 2,856 non-target trials. In this protocol, the training excerpt could be selected in a defined set, for each speaker upon three different conditions:

- **NIST-3.** The training file is the one proposed in the original NIST protocol.
- **Min.** The training excerpt is selected for a given speaker by minimizing the sum of FA and FR rates (computed on M-08).
- **Max.** The training excerpts are selected in order to maximize the sum of FA and FR rates.

Practically, the tests with the training excerpt used for a test trial are removed from the set of possible tests list after the selection of the training excerpt for Min and Max conditions.
4.2. Results

Figure 5 presents the DET curves for the three conditions (NIST-3, Min, Max). The EER are 12.1%, 4.1%, and 21.9% for NIST-3, Min, and Max conditions respectively. Looking at these results, it appears clearly that the training excerpt used for speaker model, plays an important role in the level of performance of a speaker verification system. The effect of the training excerpts cannot be neglected.

![DET curves for NIST-3 (EER=12.1%) in black, Min (EER=4.1%) in red and Max (EER=21.9%) in blue)](image_url)

5. DISCUSSION AND PERSPECTIVES

In this paper, we first analyzed the performance of a speaker verification system at the 'per speaker' level, in the same direction as [4]. The experimental framework used an open source software and a part of the 2008 NIST-SRE database. We observed differences in FA and FR rates according to the speakers behavior: 8% of the target speakers (goats) are involved in 50% of the False Reject (FR) and 26% of the speakers (lambs) are related with 50% of the False Alarm (FA). No clear boundary between the speaker types.

A second set of experiments was carried out in order to determine the role of the training excerpt. A important range of EER was observed depending on the methodology used to select the training excerpt for each speaker model. Indeed, the EER raises from 4.1% to 21.9% according to the voice sample choosen for the speaker model. This drastic sensitivity to the training material raises several questions, about the quality of the recordings, the amount of information relevant for speaker verification in the recordings and, more generally speaking, on the reliability of the results provided by the system: should we be confident on a system performance evaluation based on a global EER (or DCF) when a large part of the errors depends on the method used to select the training excerpt?

In future work, we will explore several factors like the available duration of speech in the training excerpt, the phonetic content of the training file or/and the testing file as well as the effect of the language, as possible determinant of the information relevant for speaker verification system. For example, we noticed a significant difference in terms of amount of selected speech frames (in the training excerpt) between the Min and Max condition (F=11.11, p<0.001). A preliminary study also showed some differences in terms of language mismatches: in Min, 67% of the trials have the same language in training and testing files, against 56% in Max.

6. ACKNOWLEDGEMENT

The authors thank the reviewers for their helpfull comments on this paper that should have be titled ‘A first step beyond the Doddington menagerie’.

7. REFERENCES