Two Decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology

Craig S. Greenberg[a], Lisa P. Mason[b], Seyed Omid Sadjadi[a], Douglas A. Reynolds[c]

[a] *NIST ITL/IAD/Multimodal Information Group, MD, USA*
[b] *U.S. Department of Defense, MD, USA*
[c] *MIT Lincoln Laboratory, MA, USA*

Abstract

The National Institute of Standards and Technology has been conducting Speaker Recognition Evaluations (SREs) for over 20 years. This article provides an overview of the practice of evaluating speaker recognition technology as it has evolved during this time. Focus is given to the current state of speaker recognition evaluation. Highlights from past SREs and future plans are also discussed.

Keywords: NIST SRE, Speaker Recognition, Speaker Recognition Evaluation, Speaker Verification

1. Introduction

The *Information Technology Laboratory* (ITL) at the National Institute of Standards and Technology (NIST) conducts three major activities: 1) fundamental research in mathematics, statistics, and Information Technology (IT); 2) applied IT research and development; and 3) standards development and technology transfer. Part of the ITL, the NIST Speech Group was founded in the mid-1980s to conduct these activities in service of speech-related technologies, and toward that end, held its first evaluation of automatic speech recognition technology in 1987. Since that time, the Speech Group has evolved into the Multimodal Information Group (MIG) at NIST (formerly National Bureau of Standards) and has been conducting *evaluation-driven research* of speech, text, images, video, and multimedia technologies.

Evaluation-driven research is a method of community-focused technology research that utilizes a set of common tasks, data, metrics, and measurement methods in order to reduce the total overhead necessary to conduct research and to benchmark the current state of the art and identify the most promising research directions [1]. There are four basic components that make up *evaluation-driven research*: planning, design, assessment, and a workshop. The planning component involves identifying research goals for the technology (e.g., to be able to improve performance of the fundamental underlying technology or to be robust to certain conditions), obtaining data that supports the evaluation goals, creating and documenting the evaluation plan, as well as identifying and notifying interested researchers and organizations. The design component involves deciding the tasks, metrics, and measurement methods that will make up the evaluation, and analyzing the available data to create necessary data sets (e.g., typically some data is provided to researchers in advance of the assessment period to assist in research, and other data is used as test data for the assessment). During the assessment component, either the technology developers or the evaluator runs the systems with the specified test data, and the evaluator analyzes the systems' performances. At the workshop, results and lessons learned are shared and future research goals are identified, which support the planning of future evaluations.

In 1996, NIST conducted its first evaluation of technology for automatically recognizing speakers by their voices. Over the following two decades[1], NIST conducted 15 Speaker Recognition Evaluations (SREs), in addition to an evaluation held in 2018 and evaluations planned for 2019 and 2020. During that time, speaker recognition technology has evolved substantially, and the SRE series has as well. What started as an evaluation of approximately 10 systems completing 4,000 trials has expanded into a series that commonly includes hundreds of systems completing millions of trials. This has been necessary, as the 1996 evaluation would be grossly insufficient for the research needs in 2019, and the 2019 evaluation would have been impossible in 1996–specifically, the 1996 SRE data set is too small and the data too easy to analyze performance of modern state-of-the-art systems, and the amount of data and challenging data conditions planned for SRE19 would

[1]Over the 20+ years of running the Speaker Recognition Evaluation series, NIST has received support from other U.S. government agencies, such as Department of Defense, Department of Justice and Intelligence Advanced Research Projects Activity (IARPA), to build a forum for the advancement of speaker recognition technology through *evaluation-driven research*.

have overwhelmed state of the art systems in 1996.

Despite the substantial changes the SRE series has undergone over time, certain elements have remained constant. For example, the goals of the evaluation series have always been to drive the technology forward, to benchmark the current state of the art, and to identify the most promising research directions. The evaluations have also remained open to all researchers working on the general problem of text-independent speaker recognition, and have consistently been designed to focus on core technology issues and to be simple and accessible to those wishing to participate. The requirement that submitted systems must be fully automatic and humans may not listen to, or otherwise interact with, the evaluation data has also been maintained for the entire SRE series.

In this article, we present an overview of the NIST ITL/IAD/MIG approach to evaluating speaker recognition technology over the past two decades and provide insights into what evaluations may look like moving into the next decade. The aim is to provide a review of the *evaluation-driven research* methodology employed by the SRE series that is accessible by newcomers to the field of speaker recognition evaluation. We discuss some of the key considerations necessary when conducting speaker recognition technology evaluations, and how NIST has addressed evaluating speaker recognition in general and for specific, specialized tasks. A brief survey of past SREs and results from recent evaluations is also provided, as well as a brief overview of plans for the 2019 and 2020 evaluations. We conclude the article with some general projections about how future evaluations may look as research directions have dramatically evolved since the inaugural 1996 SRE.

2. Considerations in Evaluating Speaker Recognition Technology

There is a great deal that could be said about the considerations necessary when running large-scale research-focused evaluations of speaker recognition technology.[2] Indeed, NIST has published several lengthy articles covering various aspects of this topic [2, 3, 4, 5]. While still more could be said

[2]During an informal conversation with a speech researcher, who at that time had recently worked with NIST on creating an evaluation of speaker recognition technology for the IARPA BEST program, he remarked to one of the authors that despite having been a long-time SRE participant, he was shocked by "how much actually had to be taken into account when conducting a speaker recognition evaluation."

and some material bears repeating, in the interest of focusing this article, we will limit the discussion to three main considerations: task, data, and metrics. It should be noted, however, that driving all decisions must be a set of underlying goals, framed in large part by the current maturity of the technology and the needs of the researchers, system developers, and end-users.

2.1. Task

Speakers are multifarious. Put differently, speech is a behavior, and it varies wildly both within and across individuals. As a result, speaking fixed phrases, reading, and spontaneous text-independent speech are substantially different from one another, and the performances of speaker recognition systems (and the approaches taken by these systems) in these contexts are substantially different as well [6].

Spontaneous text-independent speaker recognition has been recognized as the most general setting for speaker recognition and progress in this area seems most likely to impact other settings [2]. For this reason, NIST has chosen to make spontaneous text-independent speech the focus of the SREs. Even within this setting there are several ways of presenting the task. For example, it could be framed as an *identification task*, where the system must associate each recording with one of a fixed set of speakers (or possibly none of them); a *clustering task*, where systems must partition the speech into an unspecified number of speaker clusters; or a *detection task*, where two recordings are compared, and the task is to say whether the recordings are spoken by the same speaker[3]. An analysis of differences among various framings of the problem can be found in [2], and an argument is given in favor of detection, particularly in technology oriented evaluations. Since the goal of NIST SREs is to drive progress by focusing on the core technology, the evaluations are technology-oriented, and, as a result, the NIST SREs have been focused on spontaneous text-independent speaker detection.

While all the evaluations have had this primary task in common, several evaluations have included one or more alternate tasks. For example, speaker diarization, labeling a recording based on who spoke when, has been included in several past evaluations. This might be viewed as a segmentation task

[3]Those in the machine learning community will recognize detection as a binary classification task.

followed by a clustering task, where the recording is segmented into chunks of speech and the segments are clustered by speaker. As another example, in the 2010 and 2012 evaluations, an alternate task involved human-in-the-loop speaker recognition, also known as human assisted speaker recognition (HASR). This was a spontaneous text-independent speech speaker detection task, however humans were permitted to listen to the speech and otherwise interact with it in ways forbidden in the traditional SREs [7].

2.2. Data

"*Data is the new oil.*"[8] "*The data economy is the new economy.*"[9] While data is becoming recognized as increasingly important by society, data has always been the single most critical element of evaluation driven research. If the data is too easy, the systems will not be challenged and the evaluation is of limited value. If the data is too difficult, the systems will balk and error analysis will prove mostly fruitless. If there is not enough data, the results will lack significance. If the there is too much data[4], participants lacking the necessary compute resources will be unable to participate, the logistics of the evaluation will be burdensome, and the analysis can become impractically complex. Finally, the data must capture the desired conditions to support the specific evaluation goals and not be otherwise idiosyncratic in some detrimental way.

Past SRE data collection goals have included collection of recordings in different languages, using different microphones with varying distances from the speaker, high and low vocal efforts, noisy environments, the utilization of different communication networks and technologies, and collections with targeted speaker demographics. Originally, data was collected by offering study participants a handful of free long distance phone calls in exchange for the conversations being recorded. Due to the reduction in cost of making long distance phone calls, this model of data collection has been abandoned, instead favoring paying participants to make phone calls or be interviewed, as well as using "found" data, e.g., recordings from the internet.

Since its founding in 1992, the Linguistic Data Consortium (LDC) at the University of Pennsylvania has been the primary collector and provider of data used in the SRE series. Data collections are jointly designed by the

[4]The idea of too much data is in conflict with Bob Mercer's widely-used comment at Arden House Conference "*There's no data like more data.*" Like all general truths, there are limits to its application.

LDC and NIST, the collections are implemented by the LDC, and the data and annotations are provided to NIST. The collection is then analyzed and processed by NIST prior to splitting the data into appropriate sets for system development and evaluation. Collecting data and finding a split of the data that provides sufficient (but not excessive) amounts for system development while also allowing the necessary data for the evaluation has become increasingly difficult. The difficulty lies in the need to collect more data and that the data collected meet some specified properties. That is, precisely measuring system performance of better performing systems requires 1) more data to obtain significant results, and 2) data that is more challenging for the systems in useful ways (from a research perspective), which can prove difficult to collect.

One of the challenges of transitioning research systems into production environments is that performance "*in the lab*" varies substantially from performance "*in the field*." This has been attributed entirely to differences in the nature of the data in these two contexts. As a result, there has been an increasing move toward access to more "realistic" data in technology evaluation settings. In the SRE series, this move has recently involved the collection of telephone recordings not routed through the Philadelphia[5] public switched telephone network (PSTN), as well as including voice over internet protocol (VOIP) and audio from video (AfV) recordings. As this transition to increasingly "real" data progresses, there is a resulting loss of the carefully controlled data collection parameters, simultaneously increasing the importance and challenge of being able to measure various properties of the recordings necessary for understanding what aspects of the data are challenging for current systems. The tradeoffs can be even more nuanced. For example, selectively drawing from a real data source in a manner that eases data labeling often results in data that does not have carefully controlled independent variables and still does not sufficiently represent the data source.

2.3. Measurement & Analysis

Measurement is a foundational requirement of science and engineering. Without the ability to measure, it is not possible to distinguish between change and progress. It is difficult to overstate the fundamental importance of measurement.

[5] The LDC is located in Philadelphia, Pennsylvania, United States.

Equally important is what is being measured and how. SREs have always measured system performance using some function of error rate. This seems a bleak and arbitrary choice over focusing on success rate. However, there are advantages to focusing explicitly on errors. When the goal is to improve system performance, focusing on errors is intuitive and naturally leads to areas to direct future effort. It is also worth mentioning that the impact of halving the error rate is more apparent than a relatively small increase in success rate, which will be the case when system performance is well above chance.

As mentioned in section 2.1, the task in NIST SREs is detection, and there are two types of errors in detection tasks. Sometimes referred to as type I and type II errors in the statistics and machine learning communities, in the speaker recognition community these errors are often called misses (short for missed detections), false negatives or false rejects (when the speakers are in fact the same) and false alarms, false positives or false accepts (when the speakers are in fact not the same). Each evaluation consists of a series of trials, and a trial consists of one or more recordings of a target speaker for enrollment (or model creation) and a recording of a speaker whose identity is unknown to the system (i.e., may or may not be the target speaker) for testing purposes. Each system submitted to the evaluation must output a real-valued response for every trial, where a greater value indicates greater confidence that the enrollment and test recordings both contain speech spoken by the target speaker.

NIST has primarily measured system performance using a *detection cost function* (DCF), which is a weighted linear combination of one or more sets of false reject (aka miss) and false alarm rates observed in the evaluation trials, as the main SRE performance metric. Alternate functions over error rates have also been utilized in NIST SREs, including a function sweeping over all observable error rates [10]. Although popular among speaker recognition technology researchers due to its easy interpretability, NIST has typically not been a proponent of using the equal error rate (EER) as an SRE performance metric because of its inability to weight false alarm and false reject (miss) errors differently. NIST has found that in nearly all contexts, the applications of speaker recognition technology tend to strongly favor either few false alarms or few misses, making the equal error rate an counterproductive choice of operating point to focus attention. Instead, the SREs have focused attention on the low false positive region of the operating range, which is most appropriate for contexts where a high rate of false alarms is problematic [3], such as biometric authentication applications.

Simply measuring the performance of multiple systems on a fixed, well-chosen data set using a single, meaningful measurement is inherently valuable [11, 12]. Doing this regularly allows tracking performance progress over time. Implicit in this process is the need to understand how performance varies under different conditions present in the data, e.g., environmental noise or speaker vocal effort, as this suggests immediate research directions to improve technology performance. Analysis of SRE results have been a driver of researcher efforts as well as many data collections. Past analyses have included differences in speaker environment, vocal effort, speech modality (e.g., reading, interviews, phone conversation among strangers, phone conversations among friends), speaker aging, language, sensor, speaker demographics, and channel. NIST has also conducted analysis of the progress of speaker recognition technology over time.

As more dimensions of variation are added to the data set, more careful analysis is necessary. In order to understand how the co-occurrence of independent variables impact system performance, more data are needed, and data sets must have a sufficient number of trials to support a meaningful analysis. Further, once a relationship between an independent variable and performance has been established, a question is raised about what to do when some values of the independent variables have disproportionate representation in the evaluation data set. Recent SREs have separately measured performance across several such variables and then applied a balanced weighting to measure performance, which has also been proposed at various points in the past [13]. This approach has advantages and disadvantages, though the realized impact of this decision on SRE analysis has not been thoroughly explored.

An important, if under-recognized, aspect of analysis is how information is displayed. Numbers have relatively little meaning outside their proper context. An effective visualization method enables the interpretation process. *Detection Error Tradeoff* (DET) curves, a method that visually depicts the error rates at different operating points on a normal deviate scale, were introduced in 1997 by NIST for SRE [14]. A DET curve's general shape, distance from origin, slope, "steppiness" (or quantization), and relative distance to other DET curves are all meaningful and relatively easy to interpret, making them popular in speaker recognition as well as various other detection tasks [15, 16].

3. NIST Speaker Recognition Evaluations: A Brief History

The first SRE was held in 1996[6]. Since then, NIST has conducted more than 15 evaluations of speaker recognition technology, including a human assisted speaker recognition evaluation [7], which encouraged participation from human experts and humans collaborating with automatic systems, as well as several online challenges, which distributed embeddings to participants rather than audio recordings to reduce the barrier for participation [17]. Rather than detail each evaluation, we offer a brief summary of the early evaluations and include citations to detailed descriptions for the interested reader.

In the 1996 and 1997 evaluations, the effect of multiple-session training was explored and handset variation was featured as a prominent technical challenge. While handset variation remained a formidable challenge, the 1998 evaluation focused on matched-source training and test data [2].

The 1999 evaluation introduced two new tasks utilizing recordings with multiple speakers: multi-speaker detection, determining which speaker spoke when, and speaker tracking, performing speaker detection as a function of time [18, 19]. The test recordings for both of these tasks consisted of a recording of a telephone call mixed into a single track. The 2000 SRE (SRE00) added a speaker segmentation task, in which no specified target speakers are given and the number of different speakers may or may not be known [20]. SRE00 also included data from the Spanish *AHUMADA* corpus [21], making 2000 the first year SRE made use of non-English data.

In 2001, the SREs began including cellular data and provided automated transcripts produced by a then state-of-the-art automatic speech recognizer as part of an effort to encourage research into ideolectic features[7]. A Federal Bureau of Investigation (FBI) forensic database was included in the 2002 evaluation [23].

In 2004, NIST introduced an unsupervised adaptation mode, where the systems may optionally update the speaker model after each trial involving that model. The 2005 and 2006 evaluations [24] included recordings in

[6]NIST was involved in a limited 1992 speaker identification evaluation for a DARPA program and another small speaker identification evaluation in 1995, though it is difficult to find reference to these events elsewhere in the literature.

[7]This emphasis on higher-level features in speaker recognition was further pursued in a SuperSid workshop following the 2002 SRE [22].

multiple languages spoken by bilingual speakers as well as room microphone recordings, allowing for cross-language and cross-channel trials. This was extended in 2008 [25], by including face-to-face interview data as well. The 2010 SRE (SRE10) [26] explored several new areas, including high and low vocal effort and speaker aging, and featured a new decision cost function metric stressing even lower false positive rates. A human-assisted speaker recognition evaluation was included as part of SRE10 as well. While not part of the SRE series, in 2011 NIST conducted an evaluation of speaker recognition featuring a broad range of test conditions as part of the IARPA BEST program, most notably added noise and reverb. The 2012 SRE (SRE12) [27] explored the performance impact of allowing multiple models to be considered in a given trial by defining model speakers beforehand and distinguishing between "known" and "unknown" test speakers[8].

4. The Current State of NIST Speaker Recognition Evaluations

The 2016 Speaker Recognition Evaluation (SRE16) was not only the 20th anniversary of the SRE series, but was also the first evaluation to begin introducing a variety of changes that distinguish the current SREs from the past. These changes span all aspects of the evaluation. We highlight several of them in the contexts of evaluation administration, evaluation design, and data collection. We also offer some highlights from the most recent SREs.

4.1. Evaluation Administration

Several early SREs were impacted by delays in data collection, giving a limited amount of time to analyze, process, and organize the data sets prior to distribution[9]. This was seen as detrimental, and NIST decided to not host an SRE in 2014, which would have maintained the then biannual schedule, to allow additional time to collect and organize the data. The series resumed its biannual schedule in 2016 with SRE16.

Early SREs also included a relatively small amount of data with undesirable characteristics, e.g., a trial lacking speech, a mislabeled recording, too little data to support a more fine-grained analysis. Despite their trivial impact on performance measurement, much effort and attention went toward

[8] This turned out to be a major logistical challenge.

[9] In the 2008 SRE, the data collection finished only two weeks before the evaluation began!

dealing with these issues at the time, and they proved overly distracting, filling email threads and workshop discussions. To help limit these occurrences, NIST began collaborating with a team at MIT Lincoln Laboratory[10] to detect anomalous data and to gauge expected performance prior to the evaluation. This collaboration has been successful and has had tremendous positive impact, especially with respect to reducing data related distractions[11].

In 2016, NIST developed and began using baseline speaker recognition systems [28] to explicitly test the impact of various evaluation design decisions on system performance measurement. The use of NIST developed baseline systems has also improved NIST's ability to more precisely understand how speaker recognition technology performance has changed over time. Past evaluations have relied on researchers to voluntarily run "mothballed" systems, i.e., systems used in prior evaluations, to help assess how much a change in performance between evaluations is due to system changes and how much is due to the changes in the data. Having a collection of baseline speaker recognition systems, each utilizing the state-of-the-art approach from a past evaluation, has allowed NIST to better quantify the source of changes in performance. Additionally, evaluation participants have reported that the baseline systems' results have proven useful for debugging their research systems.

As a result of the many advances in information technology in recent years, NIST has been able to substantially improve evaluation logistics. In the past, participants needed to register for the evaluation by mail, fax, or email, and then NIST would mail them hard drives and/or optical media containing the evaluation data. Special care would be taken so that the data would be expected to arrive at all participating sites around the world at approximately the same time. The necessary logistics were burdensome and subject to human error. NIST now manages the evaluation logistics through a custom built online web platform[12], that allows sites to register for the evaluation, create formal evaluation teams composed of individual participant sites, sign all necessary documents, download data, upload system out-

[10]MIT Lincoln Laboratory also has a team that participates in the evaluations. There is no overlap in staff between these two teams and they do not collaborate on the evaluations.

[11]As performance improves, the impact of any errors in data labeling or analysis increases, further adding value to the success of this effort.

[12]After first being developed for SRE, the web platform has been used for many different technology evaluations at NIST.

put, receive the evaluation results, keys, and analysis, as well as upload and share system descriptions and workshop presentations. This change has had tremendous value for the evaluation participants as well as for NIST, substantially reducing the effort needed for, and increasing the speed of completion of, the necessary evaluation administrivia.

4.2. Evaluation Design

Prior to each evaluation, participants receive data for use in building their speaker recognition systems. It has been the common practice of SRE participants to split the provided data into training and development sets. Current evaluations have specified training and development sets within the provided data. This was in part by popular demand, but it also facilitated the introduction of *fixed* and *open* system training conditions in the evaluation series. The *fixed* training condition limits system training and development to a predetermined common set of corpora to facilitate meaningful system comparisons in terms of core speaker recognition algorithms and/or techniques. The *open* training condition allows participants to use any other proprietary and/or publicly available data in addition to the corpora provided in the fixed condition to demonstrate the gains that could be achieved with unconstrained amounts of data. Previously, training data was always unconstrained, though only data that was or would become publicly available was permitted for use.

Current SREs have also begun distributing data without speaker labels for use in system development, motivated by the availability of unlabeled data from the data source that can be useful for system adaptation. Typically, researchers have applied a clustering algorithm on this data, intending to cluster recordings based on speaker, and then model the characteristics of the various channels in the data source from the resultant clusters. Interestingly, it has been found that a perfect, or oracle, clustering of this data by speaker when using this method does not necessarily lead to optimal speaker recognition performance.

An ongoing trend in the SRE series has been the fusion of several speaker recognition systems to create a single "fusion" submission to an evaluation. While it remains interesting to see how much this approach can improve performance, there is a growing sense that the resultant fused systems complicate the error analysis and are impractical to deploy. Therefore, current evaluations have encouraged sites to also report results on their best "single" systems[13].

4.3. Evaluation Data

The data emphasis in every SRE has always been conversational telephony speech (CTS) recorded over public switched telephone networks (PSTN), though other varieties of speech data have been explored. This emphasis remains in the most recent evaluations, though two new data domains have also been introduced: voice over Internet Protocol (VOIP) and audio from video (AfV). Both the PSTN and VOIP CTS data used for the latest evaluations were extracted from *Call My Net* (CMN) 1 and 2 [29] corpora collected outside of North America, which was a new emphasis for the SREs. On the other hand, the AfV data was extracted from the *Video Annotation for Speech Technologies* (VAST) corpus [30] which was collected from amateur online video blogs (Vlogs) spoken in English, representing more modern data sources.

One factor affecting performance is the amount of speech available to the system. Current SREs explore this variability to a greater extent than in the past. It was previously common to have evaluation recordings either contain approximately 10 seconds of speech or approximately 180 or more seconds of speech for CTS data. Current evaluations now include additional segment durations spanning between 10 and 60 seconds of speech for CTS data, as well as segments potentially containing less or much more speech in the case of AfV data.

Practically speaking, recruiting subjects and collecting speech in a way that is balanced from an experimental design standpoint has always been difficult. This challenge has only grown as the number of data sources and independent variables being explored has increased. One approach is to discard data from any subject that completes only a portion of their intended recordings and then remove other subjects as well to maintain the desired balance. Large amounts of data can be discarded using this approach, so NIST has instead favored accounting for any imbalances during analysis. As mentioned in Section 2.3, current evaluations have also begun re-balancing data as part of computing the performance metric.

[13]While the definition of a “single” system is somewhat subjective, the aim is to encourage more intuitively cohesive and simplified systems versus a score level fusion of a large basket of slightly modified systems.

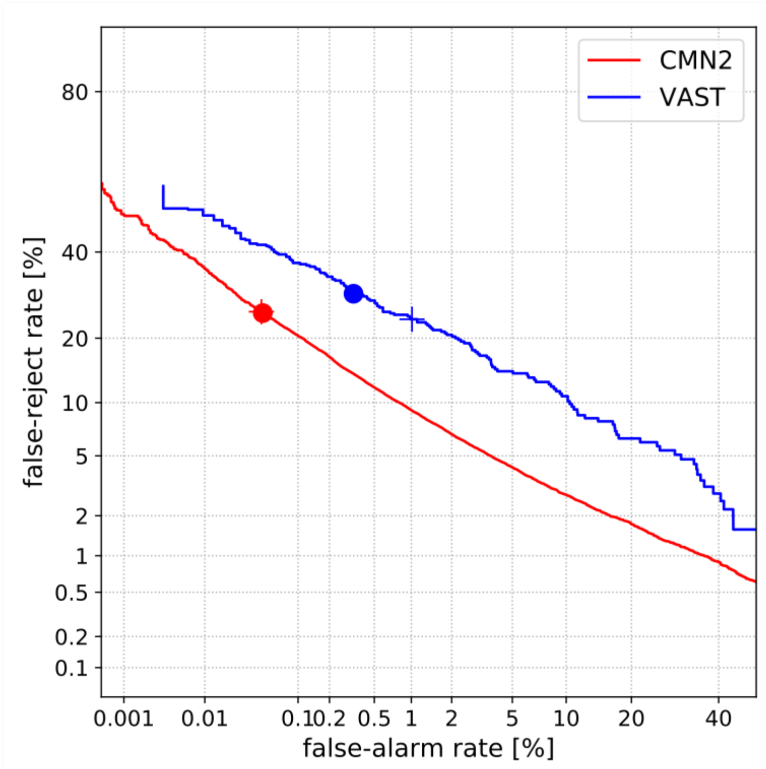

Figure 1: DET curves for a leading system's performance on CTS data (CMN2) and AfV data (VAST) in SRE18. The circles denote the operating point that minimizes the detection cost function and the cross hairs denote the operating point selected by the system. Systems performed consistently better on CTS data than AfV data in SRE18.

4.4. SRE16 & SRE18 Participation and Performance

The 2018 Speaker Recognition Evaluation (SRE18), held in September of 2018, was the latest in the series of formal NIST evaluations to support research and innovation for text-independent speaker recognition. SRE18 was organized in a manner similar to the 2016 SRE (SRE16), held in September of 2016, and included all of the above mentioned changes.

In SRE18, a total of 48 teams from 78 academic and industrial sites participated. A total of 129 valid system submissions were made, with 120 for the fixed training condition and 9 for the open training condition. The participation in SRE16 was similar, with 66 teams from 34 countries submitting 121 valid submissions (103 for fixed training condition and 18 for open training).

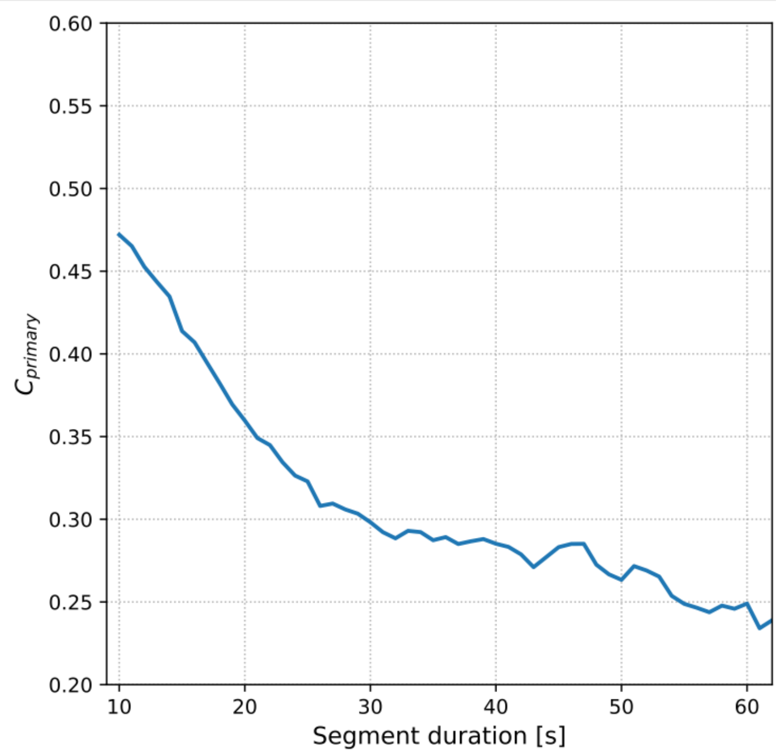

Figure 2: Performance as a function of the speech duration in a test recording for a deep learning based system submission in SRE18. Systems performed consistently better as the speech duration increased as anticipated.

These evaluations explored the impact of several factors on system performance, most notably channel/domain (Fig. 1), duration (Fig. 2), and language (Fig. 3). They also found that the effective use of the provided unlabeled development data and choice of calibration data substantially impacted system performance, particularly for the data from the AfV domain. Approaches based on recent advances in neural networks, found to be less successful in SRE16[14], were dominant in SRE18 due to the availability of large amounts of training data from a large number of speakers, the use of data augmentation in system development, and the use of more complex models.

[14]This is believed to be due to the language and domain mismatch presented in the 2016 evaluation.

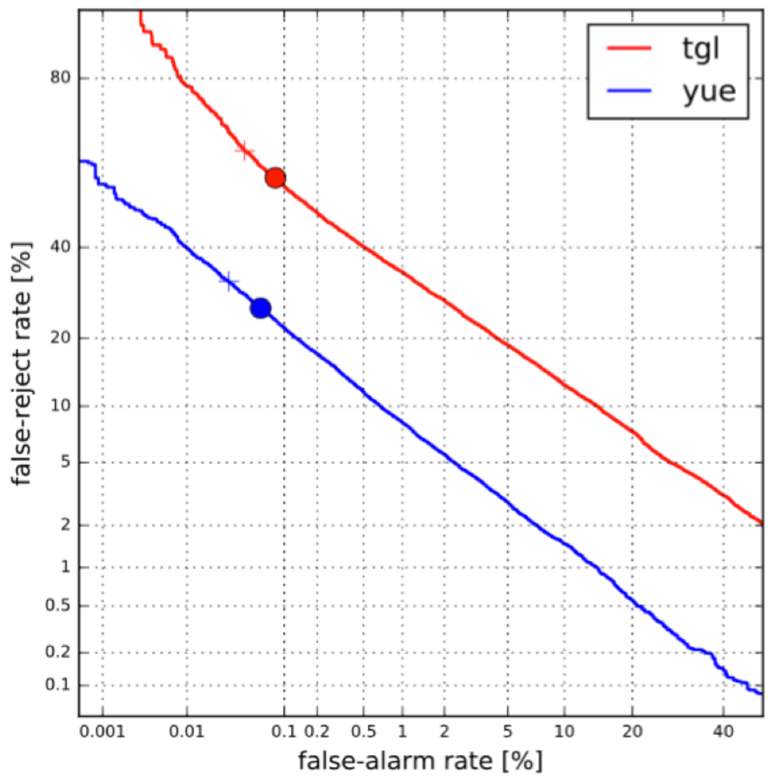

Figure 3: DET curves for a leading system's performance on Tagalog speech (tgl) and Cantonese speech (yue) in SRE16. The circles denote the operating point that minimizes the detection cost function and the cross hairs denote the operating point selected by the system. System performances were consistently better on Cantonese speech than Tagalog speech, though there were channel differences between the Tagalog and Cantonese recordings that may have led to the observed performance differences.

While fusion systems continued to maintain some of the performance advantages seen in SRE16, SRE18 witnessed strong single system results that were nearly as good as the best fused systems (Fig. 4). We include a figure comparing SRE16 systems with SRE18 systems (Fig. 5). The interested reader can find additional results for SRE16 and SRE18 in [28] and [31] respectively.

4.5. SRE19 & SRE20

Plans for the 2019 (SRE19) and 2020 (SRE20) Speaker Recognition Evaluations were publicized at the SRE18 participant workshop in December 2018. Acknowledging the observed performance challenges presented by the

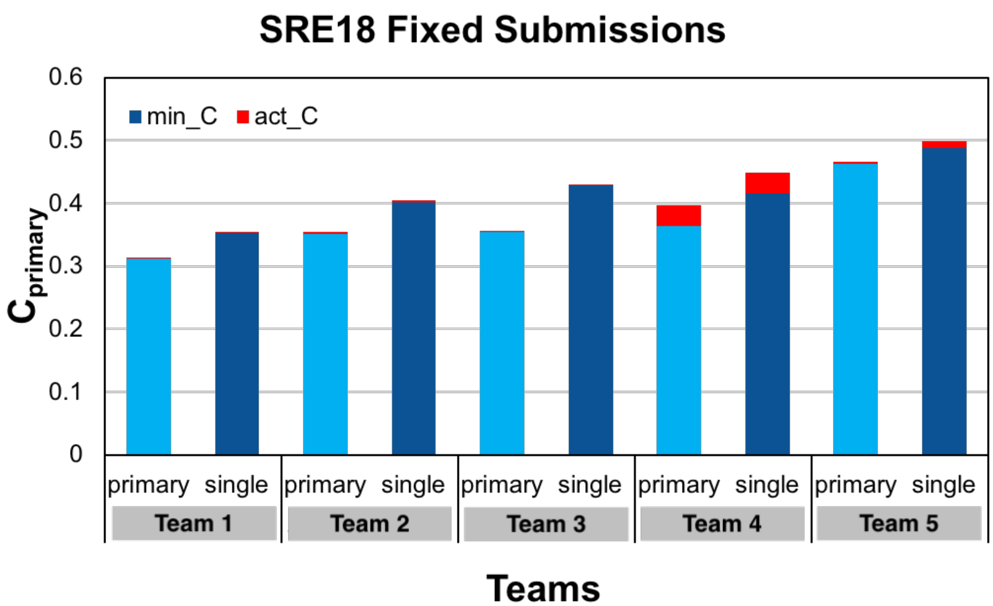

Figure 4: A comparison of system performance for fused (primary) and the best single systems from five teams in SRE18. The detection cost is displayed at both the minimum operating (min_C) and the actual operation point (act_C). The observed differences between the fused system and single systems within teams is relatively small. Further, the best single system in the evaluation was competitive with the best fused systems in the evaluation.

AfV data in SRE18 and the growing interest of the speaker recognition research community in applying speaker recognition to more realistic multimedia applications, both SRE19 and SRE20 have the goal of further exploring speaker recognition technology for audio from amateur video data. In addition to exploiting the audio from video data, these evaluations will provide participants the opportunity to explore the possibility of fusing face recognition with speaker recognition.

SRE19 will serve as a special evaluation allowing more in depth analysis and exploration into each of the data domains used in SRE18. There will be two components to SRE19: the SRE19 CTS Challenge and the SRE19 Audio-visual (AV) evaluation. The SRE19 CTS challenge will be conducted entirely online in a manner similar to the NIST 2014 and 2015 i-vector challenges [17, 32], however actual audio recordings will be used as the source data instead

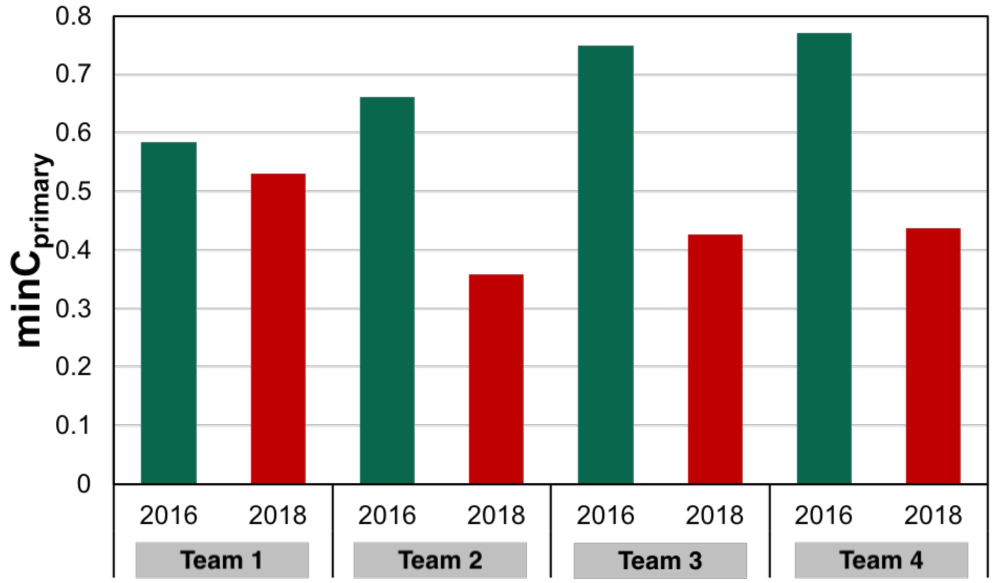

Figure 5: A comparison of system performance for the SRE16 and SRE18 systems submitted by four teams that participated in both evaluations. A data set drawn from the *Call My Net* corpus [29] was used to measure the performance of these 2016 and 2018 systems. Substantial improvements can be seen between the systems submitted in 2016 and those submitted in 2018.

of feature embeddings. Unexposed CTS data from the CMN2 corpus will be used to support the SRE19 CTS challenge. System performance scores will be made available throughout the entire evaluation period instead of at the end, and multiple submissions will be allowed, enabling participants to explore how low they can drive error rates on the traditional CTS data domain.

The SRE19 AV evaluation will be conducted in the same manner as the traditional SREs, with training and development data released in early summer 2019, evaluation data released in late summer 2019, evaluation results submitted in October 2019, and a post-evaluation workshop held in December 2019[15]. Unexposed multimedia data from the VAST corpus will be used to support the SRE19 evaluation which will feature two core evaluation tracks: audio only and audio+visual fusion. An optional visual only track will also

[15]The SRE19 workshop will be co-located with the 2019 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop in Sentosa, Singapore.

be available for participants.

The plans for SRE20 are based on the availability of a data corpus currently being collected by the LDC from multilingual speakers in both the CTS and AfV data domains. This corpus is designed to allow for explorations into cross-domain enroll-test trials (e.g. enroll on CTS data and test on AfV data for a single target speaker). The corpus is also designed to provide image data to support multimodal fusion explorations similar to SRE19. Continuing with the SRE16 and SRE18 data paradigms, this corpus is being collected outside of North America and will feature non-English data.

5. The Future of NIST Speaker Recognition Evaluations

Pending the availability of sufficient and appropriate data, it is expected that the NIST SREs will continue after 2020 and resume a bi-annual schedule in 2022 with a focus on challenging data domains and channels. NIST will also continue to explore ways to collaborate with organizers of other speaker recognition technology evaluations, where feasible, to ensure maximal community benefit. As the SRE series moves into its next decade, we highlight some of the projected trends for the future in the contexts of evaluation tasks and evaluation data.

5.1. Evaluation Tasks

The one constant throughout the SRE series from its inception has been a focus on speaker detection for spontaneous text-independent speech. The consistency of this task has allowed NIST to drive core speaker recognition technology forward and track the technological advancements over the last two decades. Moving into the next decade of speaker recognition evaluation, NIST maintains the same goal of driving speaker recognition progress by focusing on the core technology and anticipates maintaining a core focus on spontaneous text-independent speaker detection.

Continuing with the core speaker detection task will also allow NIST to have a continued focus on the technological challenges presented by data domain and channel mismatches as new domains/channels become of interest to the speaker recognition research community. And as multimedia applications become more relevant to the speaker recognition community, like realtime group discussion transcription applications that use visual data to help with speaker identification, tasks involving the fusion of audio and video data such as those introduced in SRE19 are also anticipated to continue to be considered in future SREs.

5.2. Evaluation Data

While the core SRE task will remain the same moving into the future, the data used to evaluate that task will continue to evolve in order to support exploration in more challenging domains and channels. Conversational telephony speech (CTS) data will remain a focus of the SRE series moving forward, and NIST maintains the goal of including recordings from different languages, from microphones with varying distances from the speaker, and different communication networks and technologies. It is anticipated that NIST will continue to partner with LDC to collect data for future evaluations. The collaboration has provided NIST with the largest amount of control over desired data collection parameters and data properties, which will become more important as more challenging data properties are introduced to the SRE series.

In addition to evolving CTS data characteristics, a continued progression towards data that mimics more realistic modern application conditions is also a possible focus area for future SREs (e.g., multimedia data, virtual assistant enabled devices, etc.). Recent SREs have leveraged publicly available speaker recognition data sources using "found data",[16] and this trend may continue in the future as long as these sources remain available for public research use.

6. Acknowledgements

A multitude of people worldwide have contributed to the success of the NIST SREs over the last two decades as sponsors, evaluation designers, and evaluation participants. While it is not feasible to list all their names, the authors would like to acknowledge their contributions covered in this paper.

7. Disclaimer

The results presented in this paper are not to be construed or represented as endorsements of any participants system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

[16] VoxCeleb [33, 34] and SITW [35] corpora were allowable under the SRE18 fixed training condition.

References

[1] B. J. Dorr, P. C. Fontana, C. S. Greenberg, M. Le Bras, M. Przybocki, Evaluation-driven research in data science: Leveraging cross-field methodologies, in: Proc. IEEE International Conference on Big Data (Big Data), 2016, pp. 2853–2862.

[2] G. R. Doddington, M. A. Przybocki, A. F. Martin, D. A. Reynolds, The NIST speaker recognition evaluation–Overview, methodology, systems, results, perspective, Speech Communication 31 (2000) 225–254.

[3] M. A. Przybocki, A. F. Martin, NIST Speaker Recognition Evaluation Chronicles, in: Proc. Odyssey 2004: The Speaker and Language Recognition Workshop, pp. 15–22.

[4] A. Martin, M. Przybocki, J. P. Campbell, The NIST speaker recognition evaluation program, in: Biometric Systems, Springer, 2005, pp. 241–262.

[5] M. A. Przybocki, A. F. Martin, A. N. Le, NIST Speaker Recognition Evaluation Chronicles-Part 2, in: Proc. Odyssey 2006: The Speaker and Language Recognition Workshop, pp. 1–6.

[6] A. Larcher, K. A. Lee, B. Ma, H. Li, Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication 60 (2014) 56–77.

[7] C. S. Greenberg, A. F. Martin, L. Brandschain, J. P. Campbell, C. Cieri, G. R. Doddington, J. J. Godfrey, Human Assisted Speaker Recognition in NIST SRE10, in: Proc. Odyssey 2010: The Speaker and Language Recognition Workshop, pp. 180–185.

[8] The Economist, The world's most valuable resource is no longer oil, but data, https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, 2017. Accessed: 2019-05-05.

[9] The Economist, Data is giving rise to a new economy, https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy, 2017. Accessed: 2019-05-05.

[10] N. Brümmer, J. Du Preez, Application-independent evaluation of speaker detection, Computer Speech & Language 20 (2006) 230–275.

[11] G. R. Doddington, T. B. Schalk, Speech recognition: Turning theory to practice: New ICs have brought the requisite computer power to speech technology; an evaluation of equipment shows where it stands today, IEEE Spectrum 18 (1981) 26–32.

[12] D. S. Pallett, A look at NIST's benchmark ASR tests: past, present, and future, in: Proc. IEEE ASRU Workshop 2003, pp. 483–488.

[13] D. A. v. Leeuwen, Overall performance metrics for multi-condition Speaker Recognition Evaluations, in: Proc. INTERSPEECH 2009, pp. 908–911.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, Technical Report, National Institute of Standards and Technology, Gaithersburg, MD, 1997.

[15] W. B. Croft, J. Lafferty, Language modeling for information retrieval, volume 13, Springer Science & Business Media, 2013.

[16] T. Rose, J. Fiscus, P. Over, J. Garofolo, M. Michel, The TRECVID 2008 event detection evaluation, in: Proc. IEEE Workshop on Applications of Computer Vision (WACV), 2009, pp. 1–8.

[17] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, D. A. Reynolds, The NIST 2014 Speaker Recognition i-vector Machine Learning Challenge, in: Proc. Odyssey 2014: The Speaker and Language Recognition Workshop, pp. 224–230.

[18] M. A. Przybocki, A. F. Martin, The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking, in: Proc. EUROSPEECH 1999, pp. 2215–2218.

[19] A. Martin, M. Przybocki, The NIST 1999 Speaker Recognition Evaluationan overview, Digital Signal Processing 10 (2000) 1–18.

[20] A. F. Martin, M. A. Przybocki, The NIST Speaker Recognition Evaluations: 1996-2001, in: 2001: A Speaker Odyssey-The Speaker Recognition Workshop, pp. 39–43.

[21] J. Ortega-Garcia, J. Gonzalez-Rodriguez, V. Marrero-Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification, Speech Communication 31 (2000) 255–264.

[22] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, et al., The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition, in: Proc. IEEE ICASSP 2003, volume 4, pp. IV–784.

[23] H. Nakasone, S. Beck, Forensic Automatic Speaker Recognition, in: Proc. Odyssey 2001: The Speaker and Language Recognition Workshop.

[24] M. Przybocki, A. Martin, A. Le, NIST Speaker Recognition Evaluations Utilizing the Mixer corpora – 2004, 2005, 2006, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 1951–1959.

[25] A. F. Martin, C. S. Greenberg, NIST 2008 Speaker Recognition Evaluation: Performance across telephone and room microphone channels, in: Proc. INTERSPEECH 2009, pp. 2579–2582.

[26] A. F. Martin, C. S. Greenberg, The NIST 2010 Speaker Recognition Evaluation, in: Proc. INTERSPEECH 2010, pp. 2726–2729.

[27] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, J. Hernandez-Cordero, The 2012 NIST Speaker Recognition Evaluation., in: Proc. INTERSPEECH 2013, pp. 1971–1975.

[28] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, The 2016 NIST Speaker Recognition Evaluation., in: Proc. INTERSPEECH 2017, pp. 1353–1357.

[29] K. Jones, S. M. Strassel, K. Walker, D. Graff, J. Wright, Call My Net corpus: A Multilingual Corpus for Evaluation of Speaker Recognition Technology, in: Proc. INTERSPEECH 2017, pp. 2621–2624.

[30] J. Tracey, S. Strassel, VAST: A corpus of video annotation for speech technologies, in: Proc. LREC 2018, pp. 4318–4321.

[31] S. O. Sadjadi, C. S. Greenberg, D. A. Reynolds, L. Mason, The 2018 NIST Speaker Recognition Evaluation, in: Proc. INTERSPEECH 2019, pp. 1483–1487.

[32] A. Tong, C. Greenberg, A. Martin, D. Banse, J. Howard, H. Zhao, G. Doddington, D. Garcia-Romero, A. McCree, D. Reynolds, E. Singer, J. Hernandez-Cordero, L. Mason, Summary of the 2015 NIST Language Recognition i-vector Machine Learning Challenge, in: Proc. Odyssey 2016: The Speaker and Language Recognition Workshop, pp. 297–302.

[33] A. Nagrani, J. S. Chung, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, in: Proc. INTERSPEECH 2017, pp. 2616–2620.

[34] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, in: Proc. INTERSPEECH 2018, pp. 1086–1090.

[35] M. McLaren, L. Ferrer, D. Castan, A. Lawson, The Speakers in the Wild (SITW) Speaker Recognition Database, in: Proc. INTERSPEECH 2016, pp. 812–822.