The 2016 NIST Speaker Recognition Evaluation

Seyed Omid Sadjadi[1,*], *Timothée Kheyrkhah*[1,†], *Audrey Tong*[1], *Craig Greenberg*[1], *Douglas Reynolds*[2], *Elliot Singer*[2], *Lisa Mason*[3], *Jaime Hernandez-Cordero*[3]

[1]NIST ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, Lexington, MA, USA
[3]U.S. Department of Defense, MD, USA

craig.greenberg@nist.gov

Abstract

In 2016, the National Institute of Standards and Technology (NIST) conducted the most recent in an ongoing series of speaker recognition evaluations (SRE) to foster research in robust text-independent speaker recognition, as well as measure performance of current state-of-the-art systems. Compared to previous NIST SREs, SRE16 introduced several new aspects including: an entirely online evaluation platform, a *fixed* training data condition, more variability in test segment duration (uniformly distributed between 10s and 60s), the use of non-English (Cantonese, Cebuano, Mandarin and Tagalog) conversational telephone speech (CTS) collected outside North America, and providing labeled and unlabeled development (a.k.a. validation) sets for system hyperparameter tuning and adaptation. The introduction of the new non-English CTS data made SRE16 more challenging due to domain/channel and language mismatches as compared to previous SREs. A total of 66 research organizations from industry and academia registered for SRE16, out of which 43 teams submitted 121 valid system outputs that produced scores. This paper presents an overview of the evaluation and analysis of system performance over all primary evaluation conditions. Initial results indicate that effective use of the development data was essential for the top performing systems, and that domain/channel, language, and duration mismatch had an adverse impact on system performance.

Index Terms: NIST evaluation, NIST SRE, speaker detection, speaker recognition, speaker verification

1. Introduction

NIST organized the 2016 speaker recognition evaluation (SRE) [1] in the fall of 2016. The SRE16 was the latest in the ongoing series of SRE's conducted by NIST since 1996 which serve to both stimulate and support research in robust speaker recognition as well as measure and calibrate the performance of speaker recognition systems. The basic task in the NIST SREs is speaker detection, that is, determine whether a specified target speaker is talking in a given test speech recording.

Similar to previous SREs, SRE16 focused on conversational telephone speech (CTS) recorded over a variety of handset types. However, the 2016 evaluation introduced several new aspects; first, SRE16 was run entirely online using a web platform deployed on Amazon Web Services (AWS)[1] servers.

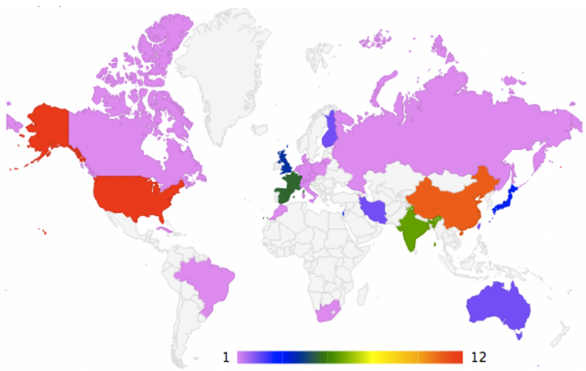

Figure 1: *Heat map of the world countries showing the number of SRE16 participating teams per country.*

The web platform supported a variety of services including evaluation registration, data distribution, system output submission, submission validation, scoring, and system description/presentation uploads. The online platform made SRE16 more readily accessible, and a total of 66 teams from 34 countries registered for SRE16. Figure 1 displays a heatmap of the number of teams per country. It should be noted that all participant information, including country, was self-reported.

Second, there were two training conditions in SRE16, namely *fixed* and *open*. In the *fixed* training condition, participants were only allowed to use *fixed* and specified data to train their systems, while in the *open* training scenario additional (publicly available) data was permitted for use in system development. System output submission for the *fixed* training condition was required for all SRE16 participants to allow better cross-system comparisons, and submission to the *open* training condition was optional but strongly encouraged to help quantify the gains that can be achieved with unconstrained amounts of data. For the 2016 evaluation, a total of 121 valid submissions were received, 103 of which were for the *fixed* training condition and the remaining 18 were for the *open* training condition.

Third, in SRE16, test segments were selected to have more duration variability than in prior evaluations. Instead of using recordings that contained nominally 20, 60, and 120 seconds of speech (such as in SRE12 [2]), the test segments were uniformly sampled, ranging approximately from 10s to 60s. This provided the opportunity to more precisely measure the impact of test segment duration on speaker recognition performance. As for speaker model enrollment, unlike previous SREs, gender labels were not provided. There were two enrollment conditions for SRE16: *one-segment*, for which systems were given

*Contractor, †Guest Researcher

[1]Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

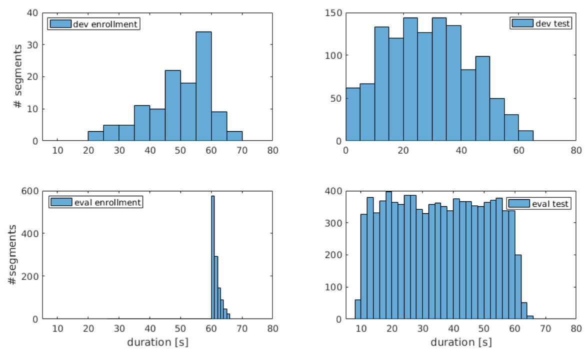

Figure 2: *Speech duration histograms for enrollment and test segments in SRE16.*

an approximately 60s long segment (in terms of active speech content as determined by speech activity detection output) to build the target speaker model, and *three-segment* where systems were given three approximately 60s long segments (all from the same phone number) to build the target speaker model. Figure 2 shows speech duration histograms of enrollment and test segments in the development and evaluations sets. It is worth noting that, similar to previous SREs, no cross-gender or cross-language trials were used in SRE16.

Fourth, unlike previous SREs, the development and evaluation sets used in SRE16 were extracted from a data corpus (i.e., Call My Net speech collection [3]) that was collected outside North America. Accordingly, the 2016 evaluation was more challenging due to the domain/channel mismatch as well as the language mismatch introduced by this dataset.

Finally, SRE16 was conducted using test segments from both same and different phone numbers as the enrollment segment(s). This was unlike most recent SREs (e.g., SRE10 [4] and SRE12 [2]) where only different phone number trials were used. The idea here was to quantify the impact of phone number match on speaker recognition performance.

2. Data

In this section we provide a brief description of the data used in SRE16 for training, development, and evaluation.

2.1. Training set

As noted previously, there were two training conditions in SRE16, namely *fixed* and *open*. In the *fixed* training condition the system training was limited to specified data sets which were as follows: i) data provided from the Call My Net corpus [3] collected by the Linguistic Data Consortium (LDC), ii) previous Mixer/SRE data [5, 6, 7], iii) Switchboard corpora (both Landline and Cellular versions) [8, 9, 10, 11, 12, 13], and iv) Fisher corpus [14]. Switchboard and Fisher corpora contain transcripts which makes them suitable for training ASR acoustic models, e.g., deep neural network (DNN) models. In addition to these, publicly available, non-speech audio and data (e.g., noise samples, room impulse responses, filters) could be used for system training and development purposes. Participation in the *fixed* training condition was required.

In the *open* training scenario, additional publicly available data was permitted for use in system development. LDC also made available selected parts from the IARPA Babel program [15] to be used in the *open* training condition. Participation in this condition was optional but strongly encouraged to help quantify the gains that one could achieve with unconstrained amounts of data.

2.2. Development and evaluation sets

In SRE16, the speech data used to construct the development and evaluation sets were extracted from the Call My Net corpus [3] collected by the LDC. The data was composed of CTS recordings collected outside North America, spoken in Tagalog and Cantonese (referred to as the *major* language), and Cebuano and Mandarin (referred to as the *minor* languages). The development set contained data from both the *major* and *minor* languages, while the test set contained data from the two *major* languages. Recruited speakers (called claque speakers) made multiple calls to people in their social network (e.g., family, friends). Claque speakers were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy home, quiet office, noisy street) for their initiated calls and were instructed to talk for 10 minutes on a topic of their choice. All segments were encoded as a-law (as opposed to mu-law used in previous SREs) sampled at 8kHz in NIST SPHERE [16] formatted files.

Participants in the 2016 evaluation received *labeled* data for development experiments that mirrored, more or less, the evaluation conditions. The development data was selected from the *minor* languages and included speech segments from 20 speakers (10 per *minor* language), and 10 calls per speaker. The participants were allowed to use the development data for any purpose (e.g., system hyperparameter tuning and adaptation).

In addition to the *labeled* development set, an *unlabeled* (i.e., no speaker id, gender, language, or phone number information) set of 2,472 calls (2,272 and 200 calls from the *major* and *minor* languages, respectively) from the Call My Net collection was made available for system training/adaptation.

3. Performance Measurement

The primary performance measure for SRE16 was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (1) specifies the primary SRE16 cost function,

$$\begin{aligned} C_{Det} = C_{Miss} \times P_{Target} \times P_{Miss|Target} \\ + C_{FA} \times \left(1 - P_{Target}\right) \times P_{FA|NonTarget} \end{aligned} \tag{1}$$

where the parameters C_{Miss} and C_{FA} are the cost of a missed detection and cost of a spurious detection, respectively, and P_{Target} is the *a priori* probability that the test segment speaker is the specified target speaker. The primary SRE16 cost metric, $C_{Primary}$, averaged a normalized version of C_{Det} calculated at two points along the detection error trade-off (DET) curve [17], with $C_{Miss} = C_{FA} = 1$, $P_{Target} = 0.01$ and $C_{Miss} = C_{FA} = 1$, $P_{Target} = 0.005$. Additional details can be found in the SRE16 evaluation plan [1].

Unlike previous SREs, in SRE16 false-reject and false-alarm counts were equalized over various partitions, where each partition was defined as a combination of: number of enrollment cuts (1-segment or 3-segment), language (Tagalog or Cantonese), gender (Male or Female), and phone number match (same or different). Furthermore, the counts were equalized over target and nontarget trials for each partition, resulting in a total of 24 ($2^4 = 16$ nontarget, 8 target) partitions [2]. More in-

[2]It is worth noting that nontarget trials from the same phone number partition were excluded.

Table 1: *Primary partitions in the SRE16 evaluation set*

Partition	# Targets	# NonTargets	# Speakers	# Calls
Male	14,960	661,652	85	595
Female	22,102	1,288,014	116	813
1conv	27,825	1,463,444	201	1408
3conv	9,237	486,222	201	1,408
Same phone#	26,024	0	197	993
Diff. phone#	11,038	1,928,594	201	1408
Tagalog	17,764	1,003,568	101	707
Cantonese	19,298	946,098	100	701

formation about the various partitions in SRE16 evaluation set can be found in Table 1. $C_{Primary}$ was calculated for each partition, and the average of the $C_{Primary}$'s for all partitions was the final metric used for system comparison.

4. Results

For each training condition (i.e., *fixed* and *open*), a team could submit up to 3 systems and designate one as the primary system for cross-team comparisons. In this section we present results for SRE16 primary submissions, in terms of minimum and actual $C_{Primary}$ as well as detection error trade-off (DET) performance curves (for more information on DET curves, see [17]).

Figure 3 shows the actual and minimum costs for all primary submissions in the *fixed* training condition. Here, the y-axis limit is set to 1 to facilitate cross-system comparisons in the lower $C_{Primary}$ region. We note that it is difficult to compare the performance of SRE16 systems to that of the prior SREs due to differences in domain/channel, language, and test segment durations. Nevertheless, compared to the most recent SREs (i.e., SRE10 [4] and SRE12 [2]), there seems to be a large drop in performance, most probably due to the noted SRE16 mismatch factors. It can be seen from the figure that, except for the top performing team, the performance gap among the top-10 teams is not remarkable. It is also observed that score calibration was successfully applied for the top performing teams (i.e., the absolute difference between the minimum and actual costs is relatively small).

Figure 4 shows system performance by training condition for the 8 teams that participated in both *fixed* and *open* tasks. We observe limited improvement in the *open* training condition over the *fixed* training condition. In some cases, worse performance is observed for the *open* training conditions, which the participants attribute to i) mismatch between the data provided for *open* training and the evaluation data, and ii) limited time and resources to effectively exploit unconstrained amounts of training data.

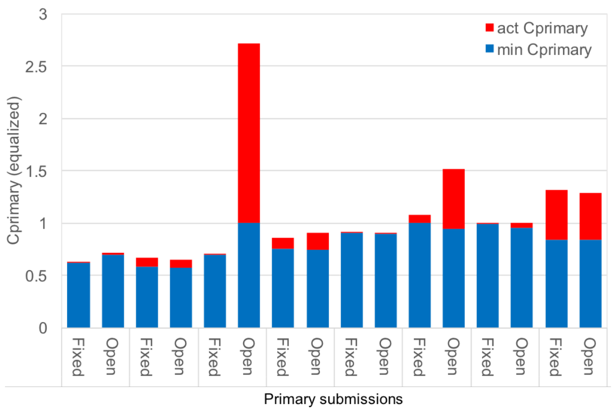

Figure 4: *Impact of open vs fixed training on performance in terms of actual and minimum $C_{Primary}$.*

Figure 5 shows DET curves for all primary submissions, with curves for top 10 systems highlighted. A similar trend is observed as in Figure 3, where, with an exception of the top performing team, the performance differences among the top-10 teams is not remarkable for a wide range of operating points.

In Figure 6 we see the DET curves for the various test segment speech durations (10s–60s). Results are shown for the top performing primary system, where filled circles and crosses represent minimum and actual costs, respectively. Limited performance difference is observed for speech durations longer than 40s. However, there is a sharp drop in performance when the speech duration decreases from 30s to 20s, and similarly from 20s to 10s. This indicates that additional speech in the test recording helps improve the performance when the test segment speech duration is relatively short (below 30 seconds), but does not make a noticeable difference when there is at least 30 seconds of speech in the test segment. It is also worth noting that the calibration error increases as the test segment duration decreases.

Figure 7 shows speaker recognition results for the top performing system as a function of language spoken in the test segment. For all operating points on the DET curves, a large performance gap is observed for Cantonese (yue) versus Tagalog (tgl). While the actual reason for such behavior remains unclear, we hypothesize that, aside from the difference in languages, the acoustic quality of the Tagalog segments as a byproduct of collection (e.g., an older telephone network) might be a contributing factor to the higher error rates for this language.

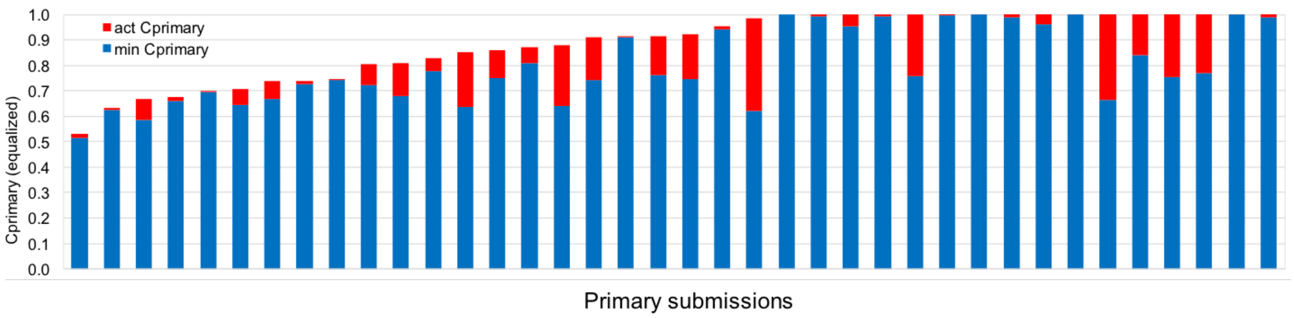

Figure 3: *Actual and minimum $C_{Primary}$ for SRE16 primary submissions.*

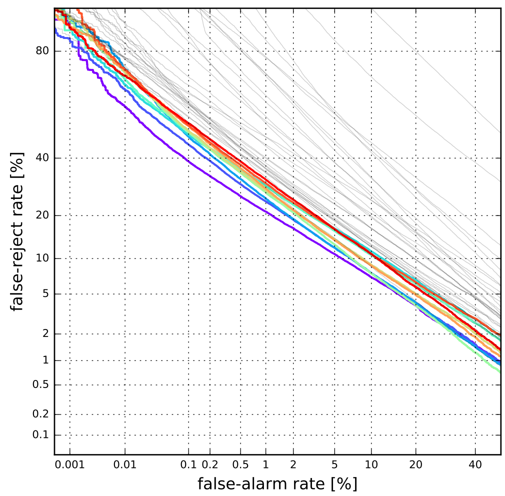

Figure 5: *DET curve performance comparison of primary submissions.*

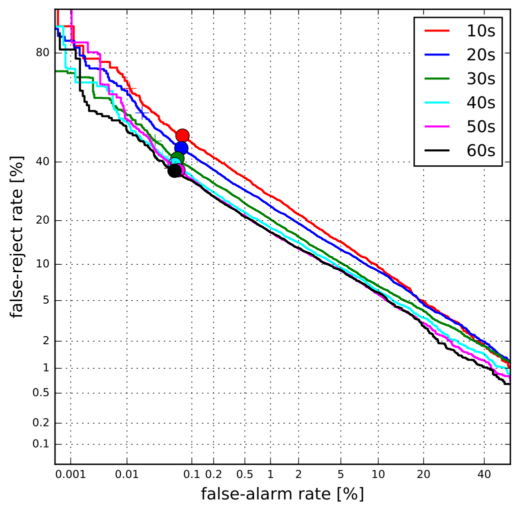

Figure 6: *DET curve performances for various test segment speech durations (10s–60s).*

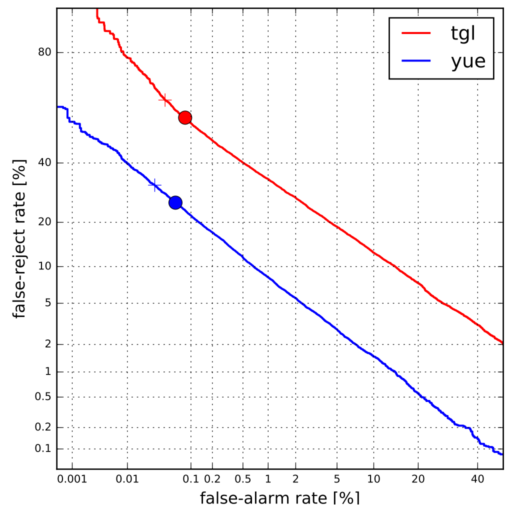

Figure 7: *DET curve performance comparison with Tagalog (tgl) vs Cantonese (yue) spoken in test segments.*

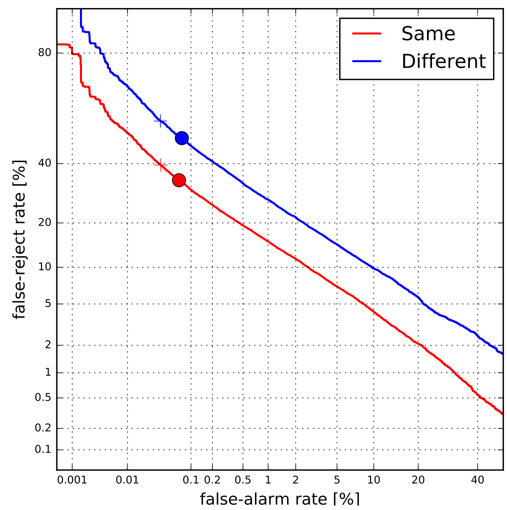

Figure 8: *DET curve performance comparison with same vs different phone numbers in enrollment and test segments.*

The impact of enrollment and test segment phone number match is shown in Figure 8. As expected, better performance is obtained when the speech segments from the same phone number are used in trials. However, the error rates still remain relatively high even for the same phone number condition. This indicates that there are factors other than the channel (phone microphone) that may adversely impact speaker recognition performance. These include both intrinsic (variations in speaker's voice) and extrinsic (variations in background acoustic environment) variabilities.

5. Conclusions

This paper presented a summary of the 2016 NIST speaker recognition evaluation whose objective was to evaluate recent advances in speaker recognition technology and to stimulate new ideas and collaborations. SRE16 introduced several new aspects, most importantly i) using *fixed* and specified training data, and ii) providing labeled and unlabeled development (a.k.a. validation) sets for system hyperparameter tuning and adaptation. There were several factors that made SRE16 more challenging than the most recent evaluations (i.e., SRE10 and SRE12), including domain/channel (due to data collected outside North America), as well as language mismatch. This motivates further research towards developing technology that can maintain performance across a wide range of operating conditions (e.g., new languages, channels, and durations).

There are plans for a follow-on analysis workshop, to be held in late 2017, as well as a new SRE, to be held during 2018.

6. Disclaimer

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

The work of MIT Lincoln Laboratory is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

7. References

[1] NIST, "NIST 2016 speaker recognition evaluation plan," https://www.nist.gov/file/325336, 2016, [Online; accessed 07-February-2017].

[2] ——, "The NIST year 2012 speaker recognition evaluation plan," https://www.nist.gov/document-6865, 2012, [Online; accessed 07-February-2017].

[3] K. Jones, S. Strassel, K. Walker, D. Graff, and J. Wright, "Call my net corpus: A multilingual corpus for evaluation of speaker recognition technology," in *Proc. INTERSPEECH (submitted)*, Stockholm, Sweden, August 2017.

[4] NIST, "The NIST year 2010 speaker recognition evaluation plan," https://www.nist.gov/document-11909, 2010, [Online; accessed 07-February-2017].

[5] C. Cieri, J. P. Campbell, H. Nakasone, D. Miller, and K. Walker, "The Mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. LREC*, Lisbon, Portugal, May 2004.

[6] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Proc. INTERSPEECH*, Antwerp, Belgium, August 2007.

[7] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. LREC*, Valletta, Malta, May 2010, pp. 2441–2444.

[8] J. Godfrey and E. Holliman, "Switchboard-1 Release 2," https://catalog.ldc.upenn.edu/LDC97S62, 1993, [Online; accessed 07-February-2017].

[9] D. Graff, A. Canavan, and G. Zipperlen, "Switchboard-2 Phase I," https://catalog.ldc.upenn.edu/LDC98S75, 1998, [Online; accessed 07-February-2017].

[10] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 Phase II," https://catalog.ldc.upenn.edu/LDC99S79, 1999, [Online; accessed 07-February-2017].

[11] D. Graff, D. Miller, and K. Walker, "Switchboard-2 Phase III," https://catalog.ldc.upenn.edu/LDC2002S06, 2002, [Online; accessed 07-February-2017].

[12] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 1 Audio," https://catalog.ldc.upenn.edu/LDC2001S13, 2001, [Online; accessed 07-February-2017].

[13] ——, "Switchboard Cellular Part 2 Audio," https://catalog.ldc.upenn.edu/LDC2004S07, 2004, [Online; accessed 07-February-2017].

[14] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC*, Lisbon, Portugal, May 2004, pp. 69–71.

[15] M. P. Harper, "Data resources to support the babel program intelligence advanced research projects activity (IARPA)," https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/harper.pdf, [Online; accessed 07-February-2017].

[16] NIST, "Speech file manipulation software (SPHERE) package version 2.7," ftp://jaguar.ncsl.nist.gov/pub/sphere-2.7-20120312-1513.tar.bz2, 2012, [Online; accessed 07-February-2017].

[17] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, Rhodes, Greece, September 1997, pp. 1899–1903.