INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The 2018 NIST Speaker Recognition Evaluation

Seyed Omid Sadjadi[1,*], *Craig Greenberg*[1], *Elliot Singer*[2], *Douglas Reynolds*[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 2018, the U.S. National Institute of Standards and Technology (NIST) conducted the most recent in an ongoing series of speaker recognition evaluations (SRE). SRE18 was organized in a similar manner to SRE16, focusing on speaker detection over conversational telephony speech (CTS) collected outside north America. SRE18 also featured several new aspects including: two new data domains, namely voice over internet protocol (VoIP) and audio extracted from *amateur* online videos (AfV), as well as a new language (Tunisian Arabic). A total of 78 organizations (forming 48 teams) from academia and industry participated in SRE18 and submitted 129 valid system outputs under *fixed* and *open* training conditions first introduced in SRE16. This paper presents an overview of the evaluation and several analyses of system performance for all primary conditions in SRE18. The evaluation results suggest that 1) speaker recognition on AfV was more challenging than on telephony data, 2) speaker representations (aka embeddings) extracted using end-to-end neural network frameworks were most effective, 3) top performing systems exhibited similar performance, and 4) greatest performance improvements were largely due to data augmentation, use of extended and more complex models for data representation, as well as effective use of the provided development sets.

Index Terms: human language technology, NIST SRE, speaker recognition, speaker verification, statistical analysis

1. Introduction

The NIST SRE18 was the latest in an ongoing series of speaker recognition technology evaluations conducted by NIST since 1996 [1], which continue to drive research and innovation in robust text-independent speaker recognition, as well as help measure and calibrate performance of state-of-the-art speaker recognition systems. SRE18 was organized entirely online using a web platform[1] that supported a variety of evaluation related services such as registration, data license agreement submission, data distribution, system output submission and validation/scoring, and system description/presentation uploads. The task in SRE18 was speaker detection, that is, determining whether a specified target speaker is talking in a given test speech recording.

SRE18 was organized in a similar manner to SRE16 [2], and offered two training conditions, *fixed* and *open*. In the *fixed* training scenario, NIST restricted system training and development data to *common* pre-specified data sets to facilitate meaningful cross-system comparisons in terms of core speaker recognition algorithms/approaches used. For the *open* training condition, participants were allowed to explore the gains that could be obtained through the utilization of unconstrained amounts of publicly available and/or proprietary data. A total of 48 teams, 21 of which were led by industrial institutions, from 78 sites made 129 valid system submissions, 120 for the *fixed* training condition and 9 for the *open* training condition. Figure 1 displays a heatmap representing the number of participating sites per country. It should be noted that all participant information, including country, was self-reported.

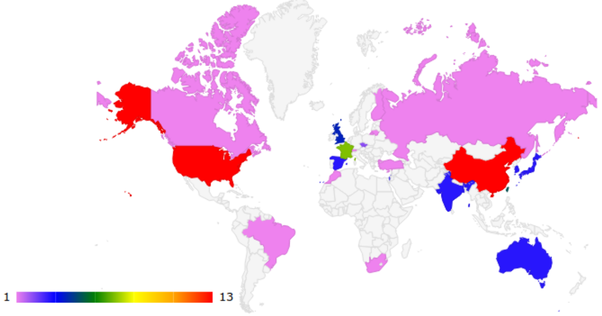


Figure 1: *Heatmap of the world countries showing the number of SRE18 participating sites per country.*

There were also a few differences between the two evaluations. In particular, SRE18 featured two new data domains; in addition to conversational speech recorded over public switched telephone networks (PSTN), VoIP data collected outside north America, as well as audio extracted from online videos (AfV) were included in SRE18 as development and test material. The PSTN and VoIP data (labeled as CTS) are spoken in Tunisian Arabic, while the AfV data is spoken in English. Unlike existing publicly available speech data derived from online "red carpet" and interview style videos featuring celebrities (e.g., VoxCeleb[2]), the AfV data in SRE18 was extracted from *amateur* online video blogs (Vlogs) that were mostly shot using personal recording devices such as cell phones in extremely diverse acoustic backgrounds. This, along with the small amount of development data, made the AfV domain more challenging than the CTS domain in SRE18.

Also, in an effort to provide reproducible state-of-the-art baselines for SRE18, NIST released well in advance of the evaluation period a report [3] containing speaker recognition system description and results obtained using both the traditional Gaussian mixture model (GMM) based as well as the recently developed deep neural network (DNN) based speaker embeddings.

[*] Contractor

[1] https://sre.nist.gov

[2] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

http://dx.doi.org/10.21437/Interspeech.2019-1351

Table 1: *Datasets for the SRE18 fixed training conditions*

Dataset	LDC Catalog ID(s)	Metadata
SRE 1996–2016	LDC2009E10 LDC2012E09 LDC2016E45 LDC2018E30 LDC2018E47	Segment/trial keys
Switchboard	LDC2018E48	ASR transcripts and segment keys
Fisher English	LDC2018E49	ASR transcripts and segment keys
MIXER 6	LDC2013S03	Segment keys
SRE 2018 DEV	LDC2018E46	Segment/trial keys
VoxCeleb[2]	–	Segment keys
SITW[3]	–	Segment/trial keys

2. Data

In this section we provide a brief description of the data used in SRE18 for training, development, and test.

2.1. Training set

As noted previously, SRE18 offered two training conditions, namely *fixed* and *open*. The *fixed* condition limited system training and development to a set of pre-specified *common* data which are listed in Table 1. SRE data from prior years (i.e., 1996–2016) along with MIXER 6 [4], Switchboard [5, 6, 7, 8, 9, 10] and Fisher [11] corpora were available from the Linguistic Data Consortium (LDC), subject to the LDC data license agreement. In addition to these, participants could use VoxCeleb[2] and SITW[3] corpora. Publicly available, non-speech audio and data, e.g., noise samples, room impulse responses (RIR), filters, could also be used, provided that a clear description was given in the final system report. Participation in the *fixed* training condition was required.

In the *open* training scenario, on the other hand, participants were allowed to utilize additional proprietary or publicly available data for system training and development. The inclusion of proprietary data was new in SRE18. Selected data from the IARPA Babel Program [12] was also made available by the LDC to be used in the *open* training condition. Participation in this condition was optional but strongly encouraged to help quantify the gains that could be achieved with unconstrained amounts of data.

2.2. Development and test sets

The speech segments in the SRE18 development (*dev*) and *test* sets were extracted from two data sets collected by the LDC to support speech technology evaluations, namely Call My Net 2 (CMN2) and Video Annotation for Speech Technology (VAST) [13] corpora. The CMN2 corpus consists of CTS recordings spoken in Tunisian Arabic, which were collected over PSTN and VoIP platforms outside north America. For CMN2 data collection, the LDC recruited a few hundred speakers called *claques* who made multiple calls to people in their social network (e.g., family, friends). Claques were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy cafe, quiet office) for their initiated calls and were instructed to talk for at least 8–10 minutes on a topic of their choice. All CMN2 recordings are encoded as a-law sampled at 8 kHz in SPHERE [14] formatted files. On the other hand, the VAST corpus contains AfV data spoken in English, which were recorded under diverse acoustic backgrounds using recording devices such as cell phones. Given the *amateur* nature of the data, each audio recording may contain speech from multiple speakers, as well as non-speech sounds such as laughter, baby crying, dog barking, etc. All VAST data are encoded as 16-bit FLAC files sampled at 44 kHz.

Table 2: *SRE18 development (dev) set and test set statistics*

Domain	Dev/Test	#speakers	#target	#non-target
CTS	Dev	25	7830	100,265
	Test	188	19,298	2,002,332
AfV	Dev	10	27	243
	Test	101	315	31,500

For system development, NIST released small development sets for both CTS and AfV data domains that broadly mirrored the test conditions. Specifically, for the CTS domain a *labeled* development set containing speech segments from 25 speakers was released for speaker enrollment and trial tests. Similar to SRE16, there were two enrollment scenarios for the CTS domain, namely 1-segment and 3-segment conditions. As the names imply, in the 1-segment condition only one approximately 60 s speech segment was given for enrollment, while in the 3-segment condition three approximately 60 s speech segments were provided to build the model of the target speaker. It is worth noting that the 3-segment condition only involved the PSTN data, because the number of VoIP calls per *claque* was limited. As part of the *dev* set for the CTS domain in SRE18, an *unlabeled* set of 2332 segments (with speech duration uniformly distributed in 10 s to 60 s range) was also made available by the LDC (LDC2018E46). The *unlabeled* segments were extracted from the non-*claque* side of the PSTN/VoIP calls. For the CTS data, the speech duration of the test segments was uniformly distributed in the 10 s to 60 s range.

As for the AfV domain, a labeled *dev* set containing audio recordings from 10 speakers was released by the LDC. The enrollment condition for the AfV domain was only 1-segment, with speech duration ranging from in 10 s to 600 s. Manually produced diarization marks also accompanied the AfV enrollment segments to facilitate building models of the primary target speaker. The AfV test segment speech duration was variable from a few seconds to several minutes, and no diarization marks were provided for the test segments.

The test sets for both domains followed exactly the same structure as described above for the *dev* set. Table 2 shows the statistics for the SRE18 *dev* and test sets.

3. Performance Measurement

Similar to the past SREs, the primary performance measure for SRE18 was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (1) specifies the SRE18 primary normalized cost function for some decision threshold θ,

$$C_{norm}\left(\theta\right) = P_{miss}\left(\theta\right) + \beta \times P_{fa}\left(\theta\right), \qquad (1)$$

[3] http://www.speech.sri.com/projects/sitw/

Table 3: *Primary partitions in the SRE18 test*

Partition	**Elements**	**#target**	**#non-target**
Gender	male	21,255	482,790
	female	39,420	1,519,542
#enrollment segments	1	48,540	1,600,871
	3	12,135	401,461
Phone# match	Y	27,456	0
	N	33,219	2,000,000
CTS type	PSTN	45,260	1,493,250
	VoIP	15,415	509,082

where β is defined as

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}}. \qquad (2)$$

The parameters C_{miss} and C_{fa} are the cost of a missed detection and cost of a false-alarm, respectively, and P_{target} is the *a priori* probability that the test segment speaker is the specified target speaker. The primary SRE18 cost metric, $C_{primary}$ was the average of normalized costs calculated at 1) two points along the detection error trade-off (DET) curve [15] for trials involving CTS data, with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.01$ and $P_{target} = 0.005$, and 2) one point along the DET curve for trials involving AfV data, with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.05$. Here, $\log(\beta)$ was applied as the detection threshold θ for computing the actual detection costs. Additional details can be found in the SRE18 evaluation plan [16].

Similar to SRE16, the CTS portion of the test data was divided into 16 partitions. Each partition is defined as a combination of: speaker gender (male vs female), number of enrollment segments (1 vs 3), enrollment-test phone number match (Yes vs No), and CTS source type (PSTN vs VoIP). However, because no actual "phone number" metadata was available for the VoIP calls, the phone number match field only contained "N" for those calls, thereby reducing the effective number of partitions to 12. More information about the various partitions in SRE18 evaluation set can be found in Table 3. $C_{primary}$ was calculated for each partition, and the final result was the average of all the partitions' $C_{primary}$'s.

4. Results and Discussion

In this section we present some key results and analyses for SRE18 primary submissions, in terms of minimum and actual costs as well as DET performance curves.

Figures 2a and 2b show performances of all primary *fixed* submissions as well as the baseline [3] system in terms of the actual and minimum costs, for the SRE18 CTS and AfV domains, respectively. Here, the y-axis limit is set to 1 to facilitate cross-system comparisons in the lower cost region. Several observations can be made from the two figures. First, performance trends on the two domains are characteristically different, with many submissions outperforming the baseline on the CTS data, but not so on the AfV data. As expected, overall detection costs and calibration errors on the relatively *cleaner* CTS domain are in general smaller than those on the AfV data which seems to be more challenging due to various factors such as loud noisy backgrounds, non-speech human vocalizations, animal sounds, diverse recording devices and codecs, and multi-party recordings, to mention a few. In addition, as noted in Section 2, the development set provided by NIST for the AfV domain was much smaller than the CTS *dev* set. Second, compared to the most recent SRE (i.e., SRE16), there seems to be a notable improvement in speaker recognition performance (see Figure 3 in [2]), which is largely attributed to the recent introduction of speaker representations (aka embeddings) extracted using end-to-end neural network frameworks [17] that can effectively exploit vast amounts of training data made available through data augmentation and/or large-scale datasets such as VoxCeleb^2. Third, it can be seen from the figures that, except for the top performing team (top two performing for the AfV), the performance gap among the top-5 teams is not remarkable. A statistical analysis of performance (e.g., confidence intervals for the cost estimates) will be reported in a future paper.

Figure 3 shows system performance by training condition (i.e., *fixed* vs *open*) for the 5 teams that participated in both conditions. We observe limited, if any, 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 used for *open* training and the evaluation data, and ii) limited time and resources to effectively exploit unconstrained

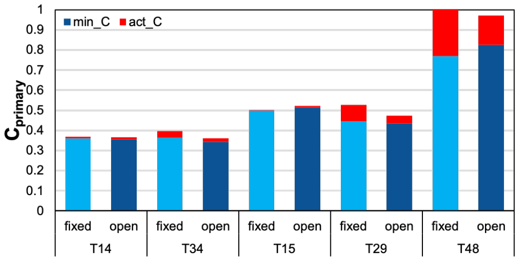
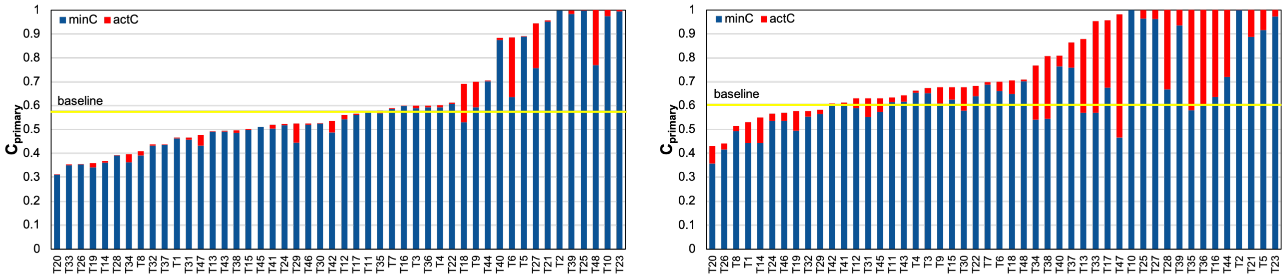


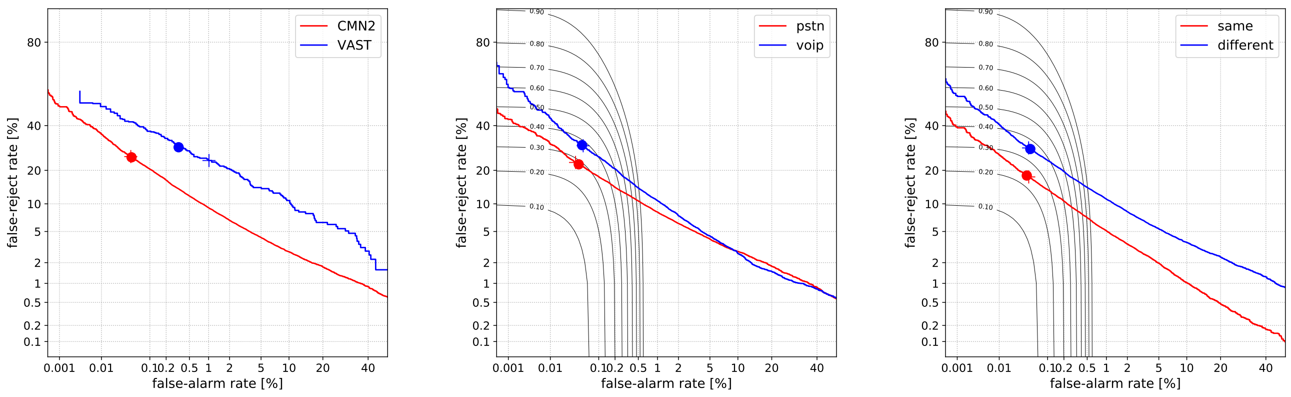
Figure 3: *Impact of open vs fixed training on performance in terms of actual and minimum costs for the SRE18 CTS domain.*



(a)

(b)

Figure 2: *Performance of SRE18 primary fixed submissions in terms of actual and minimum costs for (a) CTS, and (b) AfV domains.*



(a) (b) (c)

Figure 4: *DET performance curves for the top system by (a) data source (CMN2 vs VAST), (b) CTS type (PSTN vs VoIP), and (c) enrollment-test phone number match (same vs different). Filled circles and crosses represent minimum and actual costs, respectively.*

amounts of training data.

Figures 4a, 4b, and 4c show speaker recognition performance for the top performing submission in terms of DET curves as a function of: data source (i.e., CMN2 vs VAST), CTS type (i.e., PSTN vs VoIP), and enrollment-test phone number match for PSTN calls (same vs different), respectively. The solid black curves in Figures 4b and 4c represent equi-cost contours, meaning that all points on a given contour correspond to the same detection cost value. Firstly, consistent with our observations from Figures 2a and 2b, the detection errors (i.e., false-alarm and false-reject errors) across all operating points for the VAST domain are greater than those for the CTS domain. In addition, the calibration error for the VAST domain is much larger. Secondly, it seems from Figure 4b that for the operating points of interest (i.e., the low false-alarm region) the performance on the PSTN data is better than that on the VoIP data. We speculate this is due to: 1) VoIP being a new unseen data domain in SRE18, and 2) larger variability in devices (e.g., computers, tablets, cell phones) and accessories (e.g., wired and wireless headphones) used to make VoIP calls. Finally, as expected, better performance is observed when speech segments from the same phone number are used in trials. Nevertheless, 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.

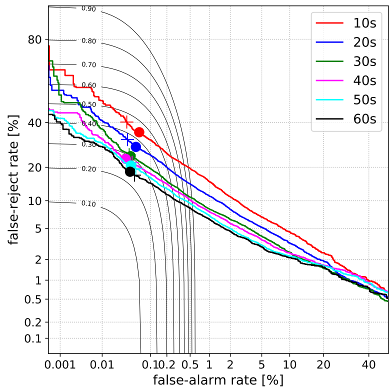


Figure 5: *DET curve performances of the top system for the various segment speech durations (10 s–60 s) in the test set.*

Figure 5 shows DET curves for the various test segment speech durations (10 s–60 s) in SRE18. Results are shown for the top performing primary *fixed* submission. Limited performance difference is observed for durations longer than 40 s. However, there is a rapid drop in performance when the speech duration decreases from 30 s to 20 s, and similarly from 20 s to 10 s. 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 (i.e., the gap between filled circles and crosses) increases as the test segment duration decreases.

5. Conclusions

We presented a summary of the NIST SRE18 whose objective was to evaluate recent advances in speaker recognition technology and to stimulate new ideas and collaborations. SRE18 featured two new data domains, namely the VoIP and the AfV, as well as a new language (Tunisian Arabic) for speaker recognition, and was the first SRE to provide an official baseline well in advance of the evaluation. Results indicate great progress in speaker recognition technology compared to SRE16, although the performance gap on the CTS domain versus the AfV domain remains relatively large. Several factors made the AfV domain more challenging than the CTS (PSTN and VoIP) domain in SRE18, including a smaller *dev* set, presence of loud background noise, animal sounds, non-speech human vocalizations, and multi-party recordings. This motivates further research towards developing a more robust technology that can maintain performance across a wide range of operating conditions (e.g., new domains, new languages, and channels).

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 Speaker Recognition Evaluation," https://www.nist.gov/itl/iad/mig/speaker-recognition, [Online; accessed 01-March-2019].

[2] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 1353–1357.

[3] S. O. Sadjadi, "NIST baseline systems for the 2018 speaker recognition evaluation," NIST, Tech. Rep., 2018.

[4] 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.

[5] J. Godfrey and E. Holliman, "Switchboard-1 Release 2," https://catalog.ldc.upenn.edu/LDC97S62, 1993, [Online; accessed 01-March-2018].

[6] D. Graff, A. Canavan, and G. Zipperlen, "Switchboard-2 Phase I," https://catalog.ldc.upenn.edu/LDC98S75, 1998, [Online; accessed 01-March-2018].

[7] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 Phase II," https://catalog.ldc.upenn.edu/LDC99S79, 1999, [Online; accessed 01-March-2018].

[8] D. Graff, D. Miller, and K. Walker, "Switchboard-2 Phase III," https://catalog.ldc.upenn.edu/LDC2002S06, 2002, [Online; accessed 01-March-2018].

[9] D. Graff, K. Walker, and D. Miller, "Switchboard Cellular Part 1 Audio," https://catalog.ldc.upenn.edu/LDC2001S13, 2001, [Online; accessed 01-March-2018].

[10] ——, "Switchboard Cellular Part 2 Audio," https://catalog.ldc.upenn.edu/LDC2004S07, 2004, [Online; accessed 01-March-2018].

[11] 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.

[12] M. P. Harper, "Data resources to support the Babel program," https://goo.gl/9aq958, [Online; accessed 01-March-2019].

[13] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[14] 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 01-March-2018].

[15] 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.

[16] NIST, "NIST 2018 Speaker Recognition Evaluation Plan," https://www.nist.gov/document/sre18evalplan2018-05-31v6pdf, 2018, [Online; accessed 01-March-2019].

[17] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 999–1003.