INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The Speakers in the Wild (SITW) Speaker Recognition Database

Mitchell McLaren[1], *Luciana Ferrer*[2], *Diego Castan*[1], *Aaron Lawson*[1]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Departamento de Computación, FCEN, Universidad de Buenos Aires and CONICET, Argentina
{mitch,aaron,dcastan}@speech.sri.com, lferrer@dc.uba.ar

Abstract

The Speakers in the Wild (SITW) speaker recognition database contains hand-annotated speech samples from open-source media for the purpose of benchmarking text-independent speaker recognition technology on single and multi-speaker audio acquired across unconstrained or "wild" conditions. The database consists of recordings of 299 speakers, with an average of eight different sessions per person. Unlike existing databases for speaker recognition, this data was not collected under controlled conditions and thus contains real noise, reverberation, intra-speaker variability and compression artifacts. These factors are often convolved in the real world, as the SITW data shows, and they make SITW a challenging database for single- and multi-speaker recognition

Index Terms: speaker recognition, database, real-world data

1. Introduction

The Speakers in the Wild (SITW) speaker recognition database offers several novel attributes to the field of speaker recognition research, including speech data from open-source media, natural speech-degrading artifacts such as noise and compression, and challenges associated with multi-speaker enrollment and test data, while consisting of enough speakers to obtain relatively narrow confidence intervals on the metrics of interest. The SITW speech data was collected from open-source media channels in which 299 well-known public figures, or persons of interest (POI), were present and speaking. Specifically, the data collection sought considerable mismatch in audio conditions, where speech for each POI was acquired both from high-quality studio-based interviews and from raw audio captured on, for example, a camcorder. Duration of speech for each speaker is unconstrained, as are the audio conditions. All noise, reverb, vocal effort, and other acoustic artifacts in the corpus are natural characteristics of the original audio. Speaking conditions include monologues, interviews, and more conversational dialogues with dominant backchannel and speaker overlap.

The SITW database was designed to offer trials (i.e., voice comparisons) involving single-speaker enrollment and test audio, as well as multi-speaker enrollment and/or test audio. Enrollment from multi-speaker audio is enabled with a small amount of ground truth annotation of where the speaker of interest speaks. This latter case is inspired by the goal of minimizing the otherwise labor-intensive task of user annotation for enrolling a speaker from a multi-speaker audio file.

The SITW database was released to the public for research purposes as part of a special session at Interspeech 2016. Accordingly, a wealth of research and analysis is present alongside this article from participants of the challenge. We encourage those interested in using the database to locate and read these articles for insight into the SITW speaker recognition challenge results and suggested research directions.

2. Existing Text-independent Databases

The field of text-independent speaker recognition benefits from large-scale evaluations such as those hosted by the National Institute of Standards in Technology (NIST). The data associated with such evaluations enables understanding how algorithms and systems perform under a discrete set of conditions. Unfortunately, many such databases are not freely available to the research community. To the best of our knowledge, several freely available, text-independent databases with more than 100 speakers are regularly evaluated in literature. These databases are focused on constrained conditions such as controlled collection of clean microphone speech [1, 2], clean speech captured from a mobile device [3], and telephone speech from forensic cases [4]. Freely available speech from databases collected for speaker diarization [5, 6] are a good resource for evaluating speaker recognition in multi-speaker audio if the same speaker appears across multiple recordings; however, processing of the data and definition of appropriate trial lists for benchmarking speaker recognition would be required. A clear lack exists in this domain for databases with a large number of speakers, real speech-degrading artifacts, and audio containing multiple speakers. Data from perhaps the widest source of video and audio, open-source media, often exhibits moderate to severe compression artifacts as well. The SITW database addresses each of these aspects in a single database, as detailed in the following sections.

3. SITW Database Description

The aim of the SITW database is to provide a large collection of real-world data with speech from individuals across a wide array of challenging acoustic and environmental conditions. Additionally, SITW includes multi-speaker audio from both professionally edited interviews (such as quiet set interviews, red-carpet interviews, and question-and-answer sessions in an auditorium) as well as more casual, conversational multi-speaker audio in which backchannel, laughter, and overlapping speech is observed. Each individual also has raw, unedited camcorder or cellphone footage in which they speak, potentially with other speakers. Importantly, all audio is naturally degraded with the noise, reverb, compression, and other artifacts included in the original audio file. The audio of the SITW database was extracted as partial excerpts of the audio track from open-source media (videos). The video was used to confirm who was speaking for the purpose of speaker labeling. An exception to this visual conformation of a speaker is in the case of radio broadcast videos in which speakers called in by telephone. Only

http://dx.doi.org/10.21437/Interspeech.2016-1129

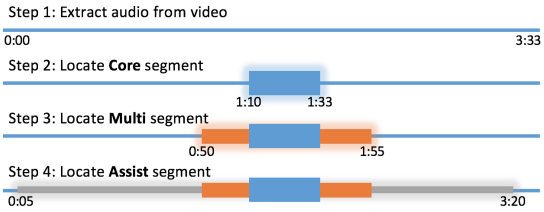

Figure 1: Up to three audio segments were defined for a given POI in a multimedia file. A core segment for the traditional single-speaker trials was defined to contain speech from just the POI. A multi segment used in multi-speaker tests was also defined where possible. Finally, an assist segment for the purpose of assisted speaker enrollment was defined using audio from the same video scene.

the extracted audio segments are released as part of the SITW database, with no option for video footage.

3.1. Audio Annotation

Figure 1 details the typical annotation process for the audio in the SITW database. It gives an example of how a single audio track from a video was segmented into several cuts for the purpose of fulfilling both the *Core* and *Assist* enrollment conditions and the *Core* and *Multi* test conditions detailed in the following sections. Note that the annotations of SITW are intended for the speaker-detection task and are not of sufficient precision in time for tasks such as measuring speaker-diarization performance.

Following Figure 1, the annotation process involved finding a continuous segment of audio in which only one POI speaks, with a goal of collecting greater than 20 seconds of speech where possible. This segment was considered the "core" segment. The audio and video around the core segment was then analyzed to define, if possible, a larger segment that contained additional speakers. This segment, when it existed, was defined as the "multi" segment. Finally, the bounds of the multi segment (or core if no multi-speaker segment was found) were extended to include almost all audio from the same scene or session while excluding overlaid music. This largest audio segment is referred to as the assist segment and is used for assisted speaker enrollment. The assist segment differs from the multi segment in that the database metadata is associated with the multi or core segment with the assist segment being found by rapid scanning of the multimedia file. In some instances, the multi segment is the same as the assist segment due to limited audio. These core, multi, and assist segments are referenced below to help describe the SITW enrollment and test conditions.

3.2. Database Meta Data

In addition to annotating audio boundaries per Section 3.1, metadata was collected to enable extended analysis of system performance[1]. The types of metadata included gender; microphone type; number of speakers; observed artifacts (noise, reverb, compression, phone); level of degradation for the most prominent artifact; recording environment; and free-text metadata associated with other observed artifacts of interest, such as laughter applause, cameras clicking, water splashing, outside noise, traffic noise, etc. While many of these labels are subjective, a single person annotated the database, which, in part,

[1]Care should be take when making subsets of trials for certain combinations of metadata, because the resulting number of speakers and trials may not allow for statistically significant comparisons.

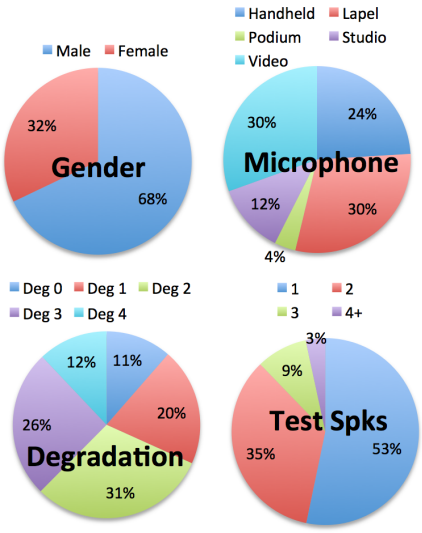

Figure 2: Data characteristics of the SITW database.

provides some consistency across such labels. The plots in Figure 2 provide a summary of how different metadata labels were observed across all audio samples in the SITW database.

3.3. Enrollment Conditions

Enrollment data for each speaker model consists of either single-speaker or multi-speaker audio, and the core or assist segment respectively from Section 3.1 as illustrated in the example in Figure 1. Approximately 5,000 speaker models exist in the SITW database. Note however that a POI refers to a unique speaker, and that several speaker models actually correspond to the same POI. For the purpose of evaluating the SITW data, all speaker models should be treated separately, as coming from different speakers, or POIs.

The two enrollment conditions are defined as:

1. ***Core***: Speaker enrollment from audio files containing a contiguous speech segment from a single POI. These audio files correspond to the core segment from the example in Figure 1 that is extracted for each POI. Conditions are unconstrained and not restricted to cleaner conditions. The amount of enrollment speech is expected to be between 6180 seconds.
2. ***Assist***[2] : Speaker enrollment from audio files that contain one or more speakers and a small annotation indicating the speaker to be enrolled. The audio from which speaker models are enrolled in this condition are the assist segments from the example in Figure 1 with a duration from 40 seconds to two hours and may contain speech from the POI of between 15 seconds to beyond an hour. Systems can utilize the provided annotation (a start and end time) along with the full audio segment to automatically locate other speech from the POI and use this speech to enroll a speaker model. Figure 3 provides an example of what might be considered enrollment speech from the assist audio segment by using semi-supervised segmentation that leverages the pro-

[2]It should be noted that for the purpose of the Interspeech 2016 SITW challenge, an additional *AssistClean* enrollment condition existed to enable participants to target cleaner enrollment conditions. Corresponding trials can be selected from the *Assist* condition by using a key provided with the database.

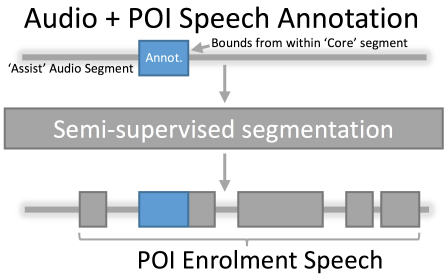

Figure 3: An example of finding enrollment speech from the assist audio segment using the provided annotation and semi-supervised segmentation.

vided annotation. The provided annotation varies between 5 seconds and 180 seconds with bias toward a shorter duration. More specifically, each original audio file is used to produce four different speaker models using annotations of 5, 10, 15, and greater than 15 seconds of speech. All annotations lie within the core segment boundaries defined in Figure 1. This provides evaluation scores for analysis of how the amount of annotation relates to system performance.

3.4. Test Conditions

While test segments may consist of acoustic conditions similar to those in the enrollment data, the degree of variation in the database is significant, and system robustness to multiple conditions will be essential to obtain a high level of performance. Two test conditions exist:

1. ***Core***: Audio files containing speech from a single speaker. The amount of speech per file is expected to range from 6 seconds to 180 seconds. As in the *Core* enrollment, audio corresponds to the core segment from the example in Figure 1.
2. ***Multi***: Audio files containing one or more speakers. This is a superset of the *Core* test condition and also includes the multi segments from the example in Figure 1. Unlike enrollment files in the *Assist* condition, no annotations or segmentation of speakers are provided for the test samples in the *Multi* condition. The amount of speech in each file will vary from approximately six seconds to ten minutes. When a POI is present in a file, the file will contain at a minimum, approximately six seconds of speech from that speaker.

4. Evaluation Protocols

The development and evaluation partitions of the SITW database were determined after the collection of the full database. This was done to ensure that no speaker "networks" existed across the partitions; that is, the POI's in the development partition never talk in the same audio file as a POI from the evaluation partition. The conditions observed in the development portion are expected to be representative of those in the evaluation portion (i.e., highly variable), as no attempt was made to emphasize mismatch between the sets.

All audio is coded as single-channel, 16-bit FLAC audio files at a 16kHz sampling rate. Note, however, that this does not necessarily reflect the bandwidth of the audio in the media uploaded to the open-source media outlet. Some audio cuts are used for multiple purposes. For example, a multi segment and assist segment from Figure 1 may be the same and therefore used in both the *Assist* enrollment condition as well as the *Multi* test condition. Additionally, the same assist segment will be used to enroll four distinct speaker models each using a different annotation for assisted enrollment. It is suggested, therefore, that the speaker model name be utilized when processing audio for the enrollment conditions rather than the audio file name, to ensure that each enrollment is handled independently (i.e., that semi-supervised segmentation with different annotations do not overwrite each other).

Table 1: The four trial conditions of the SITW database made up of two distinct enrollment conditions and two test conditions, with approximate target / impostor trial counts as summed across both the development and evaluation splits of the database.

		Test Condition	
		Core	**Multi**
Enrollment Condition	**Core**	6.3k / 1.1mil	17k / 2.9mil
	Assist	32k / 5.3mil	60k / 9.9mil

4.1. Trial Conditions

The SITW database includes four trial conditions. These are formed by the combination of the different enrollment and test conditions. For each trial condition, a list of trials is provided indicating the name of a speaker model and a test segment. Table 1 provides a matrix of the trial conditions, including the target and impostor trial counts when summed across the development and evaluation splits.

5. Factors Affecting Performance

Sourcing data from open-media has several characteristics that may affect speaker recognition performance and, in some instances, provide evaluation bias. First, in collecting speech from a well-known public figure, they often discuss similar topics either long term due to their profession (i.e., politics) or short term (i.e., an actor discussing a current movie at the time of the interview). This may provide phonetic-content overlap between audio excerpts of the same speaker that could provide advantage to those algorithms that use phonetic information. Though this may be perceived as an undesired bias, it also represents the conditions of real-world data, in which open-source media often consist of speakers associated with particular contexts.

Cross-gender trials are included in all trial conditions of SITW as these are, in numerous applications, a natural factor that automatic speaker recognition systems must cope with. Note that a bias toward male speech exists in the database, as such speech represents approximately two-thirds of the audio.

A large proportion of the SITW database was extracted from post-edited productions. This differs from most available speaker recognition databases. Quantifying the effect of this post-editing on speaker recognition performance is not the intention of the SITW database.

Speakers in multi-speaker audio that are not a POI may be present in other audio files, used either for enrollment or testing. For example, an interviewer may both be in an enrollment audio file for one POI of the *Assist* condition and also interview another POI as part of the *Multi* test condition. This also extends to the case of multiple POIs in the same multi-speaker audio file. In the *Assist* enrollment condition, this could serve to corrupt the model for speaker A with speech from speaker B, and vice versa, if robust automated semi-supervised enrollment

algorithms are not used.

Many audio segments in the SITW database were extracted from interviews. Consequently, the proportion of speech from the interview is typically swayed more toward the interviewee. In such audio segments, this may result in less perceived benefit from semi-supervised algorithms if the POI is the interviewee compared to a more speaker-balanced recording or limited presence of the POI.

The majority of the audio segments was extracted from monologues, interviews, and other multi-speaker scenarios, which naturally results in a high ratio of speech to non-speech. This does not imply that speech activity detection is not needed, but rather it should be robust to spontaneous noises that may temporarily reduce the perceived signal-to-noise ratio and reduce the amount of reliable speech information in the audio.

6. Performance Measures

Several performance measures are suggested to gauge system performance for the SITW database. A metric similar to that used in the NIST 2010 SRE formed the primary metric of the SITW speaker recognition challenge for Interspeech 2016 due to the familiarity of metric within the speaker recognition research community. Alternate metrics that aim to measure the utility of the system in the context of information retrieval, and calibration across all operating points are also suggested and recommended for evaluation. The SITW database is provided with a Python script to evaluate each of the performance metrics detailed below.

6.1. Cost Detection (C_{det})

The primary metric for SITW is based on the following detection cost function which is the same function as used in the NIST 2010 SRE, but with modified parameters. It is a weighted sum of miss and false alarm error probabilities in the form:

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}). \quad (1)$$

We assume a prior target probability, P_{tar}, of 0.01 and equal costs between misses and false alarms. The model parameters are 1.0 for both C_{miss} and C_{fa}.

For reporting, the C_{det} will be divided by the cost that a naïve system that always chooses the least costly class would get for the selected parameters. In our case, the normalization factor is given by P_{tar} [7].

For systems producing trial scores that represent calibrated log likelihood ratios, the theoretical threshold corresponding to this cost function of thr $= \log((1 - P_{tar})/P_{tar}) = 4.59$ can be applied to the scores to determine P_{miss} and P_{fa} in the calculation of C_{det}. The minimum of this value C_{det}^{min} is also a valuable metric found by sweeping a range of thresholds over the scores. This can be utilized to measure system calibration loss at the defined operating point as $C_{det} - C_{det}^{min}$.

6.2. Average R-Precision

An alternative performance measure, average R-precision ($\overline{R}_{prec}$) is calculated to indicate the "utility" of the system for retrieving the complete set of relevant test segments for a given speaker model "query". This measure is commonly used for systems concerned with information retrieval [8]. This measure accounts for the case of different numbers of relevant test segments (test segments in which the POI does speak) per query (speaker model). In short, the R_{prec} is defined as the precision in the top R scoring test segments for a given speaker model, where R is the number of target trials for that model. As an example of R_{prec}, consider a speaker model as a query for which it is known that exactly eight of the complete set of test segments for that model (as defined in the trial list) contain speech from the same POI. The R-precision of this query is then given by the proportion of the top eight scoring test segments that actually include the POI. If six of these eight contain the POI, then $R_{prec} = \frac{6}{8} = 0.75$.

It follows, then, to define the metric $\overline{R}_{prec}$ as the average R_{prec} over all speaker models for a given trial list. This average is restricted to the case where $R > 0$ such that only speaker models with target trials are included in the metric.

Given this metric, a system should aim to retrieve test segments containing a given speaker model and to place these in the highest-ranking (scoring) position for the model "query", much like the ranking of search results from an online search engine. Consequently, unlike the primary metric C_{det}, no binary detection is being made by the system and mis-calibration from one speaker model to the next model is not penalized. It is believed, however, that this measure is relevant to a real application of speaker recognition technology in which a newly acquired sample of speech forms the query to search a large database of multi-speaker audio, and these results are then parsed by an analyst. In this application, the analyst's time is most effectively used when the most relevant matches are at the top of the ranked results of the system.

6.3. Log-likelihood Ratio Cost Function

To analyze how well a system performs[3] and is calibrated across all operating points, a log-likelihood ratio cost metric, Cllr, is suggested. Assuming trials scores are represented as LLRs, then C_{llr} can be calculated as,

$$C_{llr} = \frac{1}{2 \times \log(2)} \times \left(\frac{\sum \log(1 + 1/s)}{N_{tar}} + \frac{\sum \log(1 + s)}{N_{non}} \right) \quad (2)$$

where s is the likelihood ratio for a trial, and N_{tar} and N_{non} represent the number of target and non-target trials, respectively. For more information on this metric, please refer to [9].

7. Conclusions

This article provides details on the publicly available Speakers in the Wild (SITW) database for speaker recognition research, which contains speech samples of nearly 300 well-known public figures from open-source media. This data is freely available for research purposes. All audio excerpts represent the original, "wild" audio conditions, including, for example, real noise, reverb, vocal effort, background noise, and compression artifacts. Both single- and multi-speaker audio is provided as part of the SITW corpus, enabling for multi-speaker testing as well as a speaker enrollment via a small annotation of where the speaker of interest speaks in the file. The database has already been used as part of an international speaker recognition challenge, resulting in a wealth of published research and analysis on the database. As detailed in this article, the SITW database is ideal for benchmarking the robustness of algorithms and systems to the significant variation of the audio and speaker conditions represented in open-source media.

[3] Both calibration *and* discrimination are measured with C_{llr}.

8. References

[1] G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, and D. Chow. (2015) Forensic database of voice recordings of 500+ Australian English speakers. Available: http://databases.forensic-voice-comparison.net.

[2] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The Australian national database of spoken language," in *Proc IEEE ICASSP*, 1994.

[3] C. McCool and S. Marcel, "MOBIO database for the ICPR 2010 face and speech competition," Idiap, Tech. Rep., 2009.

[4] D. Van der Vloed, J. Bouten, and D. A. van Leeuwen, "NFI-FRITS: A forensic speaker recognition database and some first experiments," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop, Joensuu, Finland*, 2014.

[5] P. Bell, M. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster *et al.*, "The MGB challenge: Evaluating multi–genre broadcast media transcription," in *Proc. IEEE ASRU*, 2015.

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *Proc. IEEE ICASSP*. IEEE, 2003.

[7] *The NIST Year 2010 Speaker Recognition Evaluation Plan*, 2010, http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf.

[8] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 1.

[9] N. Brummer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.