Corpora for the Evaluation of Robust Speaker Recognition Systems*

Douglas E. Sturim, Pedro A. Torres-Carrasquillo, Joseph P. Campbell

MIT Lincoln Laboratory
{sturim,ptorres,jpc}@ll.mit.edu

Abstract

The goal of this paper is to describe significant corpora available to support speaker recognition research and evaluation, along with details about the corpora collection and design. We describe the attributes of high-quality speaker recognition corpora. Considerations of the application, domain, and performance metrics are also discussed. Additionally, a literature survey of corpora used in speaker recognition research over the last 10 years is presented. Finally we show the most common corpora used in the research community and review them on their success in enabling meaningful speaker recognition research.

Index Terms: speaker recognition, speaker identification, speaker verification, speech corpora, speech datasets

1. Introduction

The main contribution of this paper is to aid the speaker-recognition research and evaluation communities in the selection of corpora (a substantial collection of organized speech data) and protocols on their use. The careful selection of a corpus and protocol drives the direction of research, enables the demonstration of statistically significant results, provides for comparison with other works, enables the scientific method, and improves the likelihood of acceptance of publications [1].

There is a plethora of available datasets intended for speaker recognition, including ones collected by the researchers themselves. We aim to help researchers select from among the many choices. This work expands on Campbell and Reynolds' 1999 survey [2] that ranged from TIMIT, YOHO [3] and Switchboard as used in early NIST evaluations to the SRE 2004 corpus [4], the MMSR corpora [5] [6], and the Mixer corpora [7] and presents the tremendous progress since then. This progress includes improved design of corpora, better understanding of relevant factors, protocols for the use of the data, and very large datasets.

* This work was sponsored by the Department of Justice under Air Force contract FA8702-15-D-0001. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. What makes a good Speaker Recognition Corpus?

In this section, we describe some of the desirable attributes needed for a successful and accepted speaker-recognition corpus to support research. Before discussing some of the different attributes, a couple of general comments are in order. First, any corpus developed needs to strongly consider the application(s) that it is addressing. For example, a corpus to support speaker-recognition research in forensic speaker recognition is likely very different than one for automatic speaker labeling of meetings. To support application-relevant scientific research, the corpus should support control variables and dependent variables that are relevant to the intended applications. This is not necessarily the same design as for a corpus that is designed to be predictive of performance in a given application, especially when many variables can be involved. Second, the application(s) under consideration determine the meaningful measures of performance, namely what metric(s) should be used. Good metric designs consider specific user-community needs to be informative to the research and the user communities. Furthermore, measures of effectiveness can be used to measure task performance for human consumers of speaker recognition methods. A deep discussion of metrics is beyond the scope and focus of this paper, but there are many good publications related to metrics, such as [8] [9] [10].

Given these general concepts for speaker-recognition data collections, we focus on the desirable attributes referred to earlier. These attributes include those driven by the specific effects to be studied and those driven by the nature of statistics. Factors related to statistics are those associated with the strength of the conclusions drawn from the evaluation. Ideally, these conclusions extend beyond the given evaluation and have enough predictive power for related applications, but caution must be used when making such leaps. Statistical factors include the following:

- Number of speakers. Sufficient to allow for statistically-significant performance results.
- Number of sessions per speaker. At least two sessions and sufficient to capture variabilities of the speaker, channels, conditions, and environments reflected by the application.
- Session duration. At least 30 seconds of speech time per session. Longer recordings are commonly recorded and later reduced to study duration effects.
- Intersession interval. Sessions should be recorded sufficiently separated in time to support independence assumptions and to capture variability.

- Corpus interval. The overall timespan of the corpus could be long enough to study aging effects.
- Diversity of speakers. Sufficient to capture the diversity of the speakers in the application. Balancing the demographics of the speakers permits study of the subpopulations within the evaluation sample and averages out biases from subpopulations.

An important consideration, within the context of useful statistics, is to retain in the dataset as much demographic information about the speakers as possible. This enables deeper analysis of results after the experiments are conducted by pooling specific subpopulations. For example, studying factors such as age, sex/gender, geographic origin, native language, and socioeconomic status are commonly found to be informative in subpopulation pooling experiments.

Another set of factors to be considered are those related to the effects of variation within a given speaker and across speakers. Variation is usually described in the speaker-recognition community within two areas: extrinsic variability and intrinsic variability [11] [12]. Extrinsic variability refers to factors coming from outside the speaker. Intrinsic variability refers to factors coming from within the speaker.

Extrinsic variability factors in speaker-recognition corpora commonly include the following:

- Channel. Microphone (including its placement), telephone, handset, and coding (e.g., 4G cellular, VoLTE, VoIP, wideband VTC).
- Environment. Including home/office with background noise, moving vehicle, room acoustics and reverberation. (Note that high levels of background noise, etc. can affect the speaker and cause intrinsic variability.)

Intrinsic variability factors in speaker-recognition corpora commonly include the following:

- Sex. Approximately balanced is common; although same-sex is more efficient because it does not limit cross-trials (allowing cross-sex speaker-recognition trials is generally poor practice).
- Speech style. E.g., conversational, prompted, and interview.
- Vocal effort. Including nominal, whispered, and orated.
- Languages, dialects, and colloquial speech.
- Stress. Psychological, physical, and cognitive load.
- Emotion. Including neutral, surprise, fear, and anger.
- Mental state. E.g., medicated and intoxicated.
- Physical state. Does the speaker currently have a head cold, congestion, sore throat, hoarseness, or any problems affecting the voice?

Some intrinsic variables change per session (within-subject, within-session variation) and even within a session.

Additional factors for conversational corpora include how conversation pairings/groups are determined, e.g., are the people unfamiliar, familiar, or intimate with each other and are culturally appropriate pairings used?

The factors discussed above are not all independent and they can be convolved with each other and not strictly separable. In forensic and investigatory applications of speaker recognition, many factors and variations combine and add up to extreme situational mismatches between the voice samples to be compared. Designing and collecting corpora to reflect this is a considerable challenge.

3. Corpora in Speaker Recognition Publications

In this section, we analyze what speech corpora the speaker recognition research community has conducted studies on during the past 10 years. In order to sample the community, we considered the annual IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) sessions on speaker recognition and speaker verification over the years 2006 through 2016. We chose the historical ICASSP conferences because of the consistency of the sessions in the area of speaker recognition. Papers from all of the sessions on speaker recognition/verification were reviewed. The number of sessions in speaker recognition each year varied from 3 to 5 over this period.

All speech corpora used in each paper were tabulated. A heat map of the occurrences of the top corpora is presented in Table 1. For a speech corpus to be counted as an occurrence, it must be used in training or evaluation of the research system presented in each paper. The table only displays the top 12 most popular corpora used by the research community in the development or evaluation of speaker recognition systems.

Table 1: Heat-map of corpora appearing in the ICASSP sessions on Speaker Recognition years 2006-15.

Year	Own	RATS	RSR15	SRE12	SRE10	SRE08	SRE06	SRE05	SRE04	PRISM	SRE03	YOHO	TIMIT
2006	3	0	0	0	0	0	0	5	6	0	2	0	1
2007	6	0	0	0	0	0	8	10	9	0	1	0	1
2008	1	0	0	0	0	8	17	12	12	0	1	0	0
2009	2	0	0	0	0	9	17	13	13	0	1	0	1
2010	3	0	0	0	0	7	2	0	0	0	0	2	0
2011	5	0	0	0	17	7	4	0	2	0	2	0	4
2012	2	0	0	1	7	3	1	0	2	0	1	1	1
2013	5	4	1	5	9	4	0	0	0	1	0	0	1
2014	4	1	2	6	11	4	1	2	0	2	0	2	0
2015	3	3	3	2	4	4	5	5	4	0	0	0	0

Descriptions of the top 12 corpora seen are presented in section 3.2.

Research papers that use self-recorded speech data were tabulated and denoted as the "*Own*" column in Table 1. The problems with using private self-recorded speech data are that it does not allow repeatable scientific experiments within the research community or performance comparisons. The premise of the NIST speaker recognition evaluations (NIST-SRE) was for the research community to build on each other's results and move the science forward.

Table 1 and Table 2 display some distinct trends in community research in speaker recognition over the past 10 years. The volume of speaker recognition research trends upwards in the years following a NIST-SRE. One of the goals of the speaker recognition evaluations is to stimulate the community in conducting research. The SRE seems to fulfill this purpose.

Another trend in our review is that speaker recognition systems now use more speech corpora in development. Research sites frequently use most of the speech data they have available. This tends to make the systems more robust to NIST-SREs, but also requires more computing resources to field a competitive system. A new site trying to enter the research community may have difficulty in reproducing previous work. To combat this, the community has posed challenge problems. An example is the NIST-SRE i-vector Machine Learning Challenge [13], where the common preprocessing of the speech is provided for research sites.

An exception to the above trend is the lukewarm response the research community had to the NIST 2012 Speaker Recognition Evaluation. The research community's publications decreased after most recent NIST-SRE 2012.

Table 1: Number of sessions in speaker recognition/ verification at ICASSP (* indicates evaluation year).

Year	Sessions/Year
2006*	3
2007	4
2008*	3
2009	3
2010*	3
2011	5
2012*	3
2013	3
2014	4
2015	3

3.1. Metrics and Protocols

How a corpus is evaluated is just as important as the characteristics of the recorded speech. Metrics and protocols are defined by the evaluation. Typical protocols can take the form: 1) defined set trials between training/testing utterances and 2) specific restrictions on training of hyperparameters and score calibrators.

Performance metrics provide a meaningful measure of system performance. Standardized metrics allow ease of comparison to other researchers in the community. These can take the form of 1) equal error rate (EER), 2) detection error tradeoff (DET) curve [14], 3) calibrated log-likelihood ratio (CLLR) [9], and 4) detection cost function (DCF) [14].

For the purpose of this paper, we are limiting the metrics and protocols to those frequently seen in the past ten ICASSP conferences. The most frequent protocols were the following:

- NIST-SRE 1996-2010 and 2016 (SRE)
- NIST-SRE 2012 (SRE-2012)
- Robust Speaker Recognition 2015 (RSR2015)
- Site specific protocol (Own)

We define the evaluation protocol to be SRE or SRE-2012. Before the 2012 NIST-SRE evaluation, protocol was based on independent trials. A system decision is based only on the specified test segment and target speaker model. This protocol temporarily changed in 2012 to where each detection decision could now use knowledge of models and segments in other trials.

Table 3 shows a heat map of the protocols used in the past 10 years of ICASSP sessions on Speaker Recognition. The SRE protocol dominates the literature. This is followed by self-defined protocols. Unfortunately self-defined protocols make it hard for the community to compare research. As seen in the previous section, the SRE-2012 protocol has not been widely adopted by the research community. Thankfully, the SRE-2016 protocol has gone back to independent trials [15].

Table 2: Heat map of protocols used in ICASSP sessions on Speaker Recognition in years 2006–15.

Year	Own	RSR2015	SRE	SRE-2012
2006	6	0	12	0
2007	14	0	12	0
2008	5	0	19	0
2009	2	0	19	0
2010	14	0	10	0
2011	14	0	31	0
2012	7	0	13	1
2013	7	0	17	5
2014	3	0	19	6
2015	7	5	6	2

3.2. Corpora Descriptions

NIST Speaker Recognition Evaluations 1996–2012

The vast majority of speaker recognition corpora available to the speaker ID community over the years have been through speaker recognition evaluations conducted by the National Institute of Standards and Technology (NIST).

We also noted that the evaluations drew from LDC collections of Switchboard 1 and 2, Mixer, and Fisher. These corpora were repackaged by NIST and distributed for the SREs. The SRE corpora and protocols are the community standard of reporting results.

The NIST-SREs have predominantly been on telephone channels. The extrinsic factors for the telephone tasks are handset changes and channel encoding, as communications has moved away from landline to cellular. Starting in 2004, the room microphone condition was introduced. This yielded a larger variety of extrinsic factors, such as microphone types and placements, room geometries and acoustics, and background noises.

Intrinsic variabilities include session, language, speech style, and vocal effort. The high-vocal effort tasks are limited. The majority of the tasks are centered on polite conversations between strangers.

TIMIT and Derivatives (LDC)

The TIMIT corpus was released in 1993 and is one of the first publicly available speech datasets. It was intended for development and evaluation of automatic speech recognition systems. This is a weak speaker-recognition corpus by today's standards. The intrinsic and extrinsic variability is very limited, since TIMIT has no intersession variability and unrealistically pristine recording conditions. The TIMIT corpus poses little challenge for speaker recognition research.

Robust Automatic Transcription of Speech (RATS)

The goal of the RATS corpus [16] was to provide a challenging channel for the community to conduct speech research. The data was generated by rebroadcasting previously recorded telephone speech from NIST-SRE, NIST language recognition evaluations, and Voice of America telephone call-ins. The data supports research in speaker recognition, language identification, speech activity detection, and other topics. The dataset contains the speech from Arabic, Farsi, Pashto, Urdu, and English speakers.

Since the speech is a rebroadcast of telephone data, the intrinsic variabilities include session, language, and speech style. The extrinsic variability now additionally includes degraded radio channels.

Robust Speaker Recognition 2015 (RSR2015)

The Robust Speaker Recognition 2015 corpus [17] supports text-dependent speaker recognition. It centers on humans interacting with computers with a constrained set of pass-phrases, commands, and digit strings. The intrinsic variabilities are limited in the text-dependent tasks over multiple sessions. The extrinsic variability of recording microphones is limited to the six portable devices from different manufacturers. Background room effects add to the extrinsic variability of the recordings.

Promoting Robustness in Speaker Modeling (PRISM)

The PRISM corpus [18] is a meta-corpus, reorganizing the previously exposed NIST-SRE speech corpora into a single speaker recognition database. The effort was to agglomerate and document all of the variabilities seen in the exposed speech data of NIST-SREs from 1996–2010. PRISM provides a nice database organization of the past NIST-SREs. Since it utilizes SRE data, the intrinsic and extrinsic variabilities are identical to the SREs.

YOHO

The YOHO speech data corpus [3] was collected to support text-dependent speaker authentication research. The application of the research was for access control to a secure telephone unit (STU). Speech was recoded in a quiet office environment with the STU's high-quality microphone. The clean recording conditions greatly limit the extrinsic variability of the data.

Recorded subjects were prompted to say randomly generated triple two-digit numbers simulating a lock combination. This limits the intrinsic variation of speaking style. The lock combination vocabulary space allows for expansion without requiring all entire lock combination phrases to be seen in training, which is not the usual interpretation of text-dependent speaker recognition.

Table 3 shows additional attributes from three selected corpora. These were selected as an interesting contrast of what the speaker recognition community chose to do research on.

Table 3: Attributes of selected corpora most used in research.

Attribute	SRE-2010	RATS	RSR2015
# of speakers	499 (239 M/260 F)	710	298(142 M/156 F)
# sessions/speaker	average 11	10	9
Intersession interval	weeks	weeks	weeks
Type of speech	conversation	conversation	fixed speech
Microphones	telephone / room mic	telephone	mobile device mic
Channels	handset	radio rebroadcast	office
Acoustic environment	telephone-locations / office	telephone-locations	telephone-locations
Evaluation procedure	Yes [19]	Yes [20] [16]	Yes [17]

4. New Resources

At the time of publication, new corpora and catalogs became available to consider. For example, the Speakers in the Wild (SITW) dataset is new and it supports an Interspeech 2016 Challenge [21]. The RedDots dataset supports another Interspeech 2016 Challenge with the "goal of exploring new directions and better understanding of speaker-channel-phonetic variability modeling for text-dependent and text-prompted speaker verification over short utterances" [22]. An especially valuable resource is the extensive dataset catalog created by the ISCA SIG on Robust Speech Processing (RoSP) [23]. An interesting corpus with concurrent speech and MRI recordings is also available and enables a unique understanding in human speech production [24] [25].

5. Conclusions and Recommendations

We described significant corpora available to support speaker recognition research and evaluation, along with details about the corpora collections and designs. In case we missed any important corpora and protocols, please contact the authors. We also described the attributes of high-quality speaker recognition corpora. Considerations of the application, domain, and performance metrics are also discussed. Additionally, a literature survey of corpora used in speaker recognition research over the last 10 years is presented. Finally, we show the most common corpora used in the research community and review them on their success in enabling meaningful speaker recognition research. We recommend the careful selection of a corpus and protocol to focus your research, enable the demonstration of statistically significant results, provide for comparison with other works, enable the scientific method, and improve the likelihood of acceptance of publications.

6. Acknowledgements

We thank the NIST Organization of Scientific Area Committees' (OSAC) Speaker Recognition Subcommittee for their contributions to this work.

7. Bibliography

[1] J. R. Matey, G. W. Quinn, P. Grother, E. Tabassi, C. Watson and J. Wayman, "Modest proposals for improving biometric recognition papers," in *The IEEE Seventh International Conference on Biometrics: Theory, Applications and Systems*, Arlington, Sep 2015.

[2] J. P. Campbell and D. A. Reynolds, "Corpora for the Evaluation of Speaker Recognition Systems," in *Proc. International Conference on Acoustics, Speech, and Signal Processing in Phoenix*, Arizona, 1999.

[3] J. P. Campbell, "Testing with the YOHO CD-ROM Voice Verification Corpus," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Detroit, May 1995.

[4] A. F. Martin, D. Miller, M. A. Przybocki, J. P.Campbell and H. Nakasone, "Conversational Telephone Speech Corpus Collection for the NIST Speaker Recognition Evaluation 2004," in *Proc. 4th International Conference on Language Resources and Evaluation*, Lisbon, May 2004.

[5] J. P. Campbell, H. Nakasone, C. Cieri, D. Miler, K. Walker, A. F. Martin and M. A. Przybocki, "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Toledo, 2004.

[6] C. Cieri, J. P. Campbell, H. Nakasone, K. Walker and D. Miller, "The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data," in *4th International Conference on Language Resources and Evaluation* , Portugal, 2004.

[7] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki and K. Walker, "The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research," in *International Conference on Language Resources and Evaluation (LREC)*, Genoa, May 2006.

[8] D. V. Leeuwen and N. Brummer, "An Introduction to Application-Independent Evaluation of Speaker Recognition System," in *In Speaker Classification I*, vol. 4343, The series Lecture Notes in Computer Science, 2007, pp. 330-353.

[9] N. Brummer, "Application-independent evaluation of speaker detection," in *Proc. Odyssey, Speaker and Language recognition workshop*, 2004.

[10] N. Brummer and J. D. Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20.2, pp. 230-275, 2006.

[11] M. Graciarena, T. Bocklet, E. Shriberg, A. Stolcke and S. Kajarekar, "Feature-Based and Channel-Based Analyses of Intrinsic," in *INTERSPEECH*, 2009.

[12] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74-99, 2015.

[13] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET curve in assessment of detection task performance.," in *Proc. Eurospeech*, Rhodes, Greece,, 1997.

[15] National Institute of Standards and Technology, "2016 NIST Speaker Recognition Evaluation," NIST, 2016.

[16] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey*, 2012.

[17] A. Larcher, K. A. Lee, B. Ma and H. Li, "RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *Interspeech*, 2012.

[18] L. Ferrer, H. B. L. Bratt, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of NIST 2011 workshop*, 2011.

[19] National Institute of Standards and Technology, "The NIST Year 2010 Speaker Recognition," NIST, 2010.

[20] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S. Mallidi, N. Mesgarani and R. Schwartz, "Developing a speaker identification system for the DARPA RATS project," in *ICASSP*, 2013.

[21] M. McLaren, L. Ferrer, D. Castan and A. Lawson, "Speakers in the Wild (SITW) Speaker Recognition Database," in *Interspeech 2016*.

[22] K. A. Lee, A. Larcher, H. Aronowitz, G. Wang and P. Kenny, "The RedDots Challenge: Towards Characterizing Speakers from Short Utterances," 2016. [Online]. Available: https://sites.google.com/site/thereddotsproject/reddots-challenge.

[23] ISCA Special Interest Group on Robust Speech Processing, "Datasets – RoSP," [Online]. Available: https://wiki.inria.fr/rosp/Datasets.

[24] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307-1311, 2014.

[25] E. Godoy, A. Dumas, J. Melot, N. Malyska and T. F. Quatieri, "Relating estimated cyclic spectral peak frequency to measured epilarynx length using Magnetic Resonance Imaging," in *Submitted to Interspeech*, 2016.

[26] National Institute of Standards and Technology, "The NIST Year 2012 Speaker Recognition," NIST, 2012.