CORPORA FOR THE EVALUATION OF SPEAKER RECOGNITION SYSTEMS

Joseph P. Campbell, Jr.[1] *and Douglas A. Reynolds*[1,2]

[1]Department of Defense
9800 Savage Rd, Ste 6516
Ft. Meade, MD 20755-6516, USA
j.campbell@ieee.org

[2]MIT Lincoln Laboratory
244 Wood St
Lexington, MA 02420-9185, USA
dar@sst.ll.mit.edu

ABSTRACT

Using standard speech corpora for development and evaluation has proven to be very valuable in promoting progress in speech and speaker recognition research. In this paper, we present an overview of current publicly available corpora intended for speaker recognition research and evaluation. We outline the corpora's salient features with respect to their suitability for conducting speaker recognition experiments and evaluations. Links to these corpora, and to new corpora, will appear on the web http://www.apl.jhu.edu/Classes/Notes/Campbell/SpkrRec/. We hope to increase the awareness and use of these standard corpora and corresponding evaluation procedures throughout the speaker recognition community.

1. INTRODUCTION

The use of standard speech corpora for development and evaluation is one of the major factors behind progress over the last 10 years in automatic speech processing research, particularly in speech and speaker recognition. Perhaps the main benefit of using standard corpora is that it allows researchers to compare performance of different techniques on common data, thus making it easier to determine which approaches are most promising to pursue. In addition, standard corpora also can be used to measure current state-of-the-art performance in research areas for particular tasks and highlight deficiencies that require further research.

In this paper, we present a survey of standard speech corpora that are useful for development and evaluation of speaker recognition systems for various application tasks. The corpora listed here were selected based on public availability and applicability for evaluating speaker recognition systems. In particular, we focused on four factors: 1) number and diversity of speakers; 2) number and time separation of sessions per speaker; 3) type of speech (e.g., fixed-phrase, prompted digits, read sentences, conversational speech); and 4) channel, microphone, and recording environment types and variability (e.g., wideband microphone in sound booth, variable telephone handsets in home/office environment). The degree to which a corpus exhibits these factors determines its utility as a valid evaluation/development corpus. Using a corpus to experiment on or evaluate a speaker recognition system requires the definition of an evaluation procedure that specifies, among other things, the partitioning of a corpus into training and testing data sets. Examination of system performance for specific conditions (e.g., performance when training and testing speech are recorded with different microphones) requires that the corpus have enough speech from enough speakers for the condition of interest to construct a valid experiment. In addition, for speaker verification experiments, a corpus may need to be large enough to specify a development set of speakers so that not all speakers are used in training system parameters, allowing for realistic, unseen imposter test speech.

While important, there is insufficient space here to provide a detailed discussion of how to design valid evaluation procedures for the corpora. The reader is referred to [1] for an overview of speaker recognition evaluation methodology and to [2] for examples of speaker recognition evaluation plans. We will note when listed corpora have predefined evaluation procedures.

2. AVAILABLE CORPORA

In this section, we present a survey of publicly available corpora for speaker recognition. The primary suppliers of these corpora are the European Language Resources Association (ELRA) [3], the Linguistic Data Consortium (LDC) [4], and the Oregon Graduate Institute (OGI) [5], as indicated next to each corpora name. Please contact these organizations for information on acquiring the corpora.

2.1 TIMIT and Derivatives (LDC)

The TIMIT corpus was designed to provide speech data for acoustic-phonetic studies and for developing and evaluating automatic speech recognition systems. Since it was one of the first corpora available with a large number of speakers, it has been used for many speaker recognition studies. TIMIT and its derivatives, however, are poorly suited for evaluating speaker recognition systems primarily due to the unrealistically pristine conditions (no intersession variability and wideband recordings in a sound booth). The TIMIT family of corpora are useful to some extent for contrastive-type experiments to attempt to isolate and quantify the effect of specific degradations imposed on pristine data [6].

TIMIT related corpora include: 1) **FFMTIMIT** (recordings from the original TIMIT recording sessions made from a far field secondary microphone); 2) **NTIMIT** (created by NYNEX by playing TIMIT speech through an artificial mouth into a carbon-button telephone handset, transmitting the speech over local and long-distance telephone lines and recording the received signal; 3) **CTIMIT** (created by Lockheed-Sanders by playing TIMIT speech into a cellular telephone handset in a moving van, transmitting over the cellular network and recording the signal at a central location; and 4) **HTIMIT** (created by Lincoln Laboratory by playing TIMIT speech through various electret and carbon-button telephone handsets and recording the signal directly from the handset output).

Table 1 TIMIT Corpus Description

# of speakers	630 (438 M/192 F)
# sessions/speaker	1
Intersession interval	none
Type of speech	Read sentences
Microphones	Fixed wideband headset
Channels	Wideband/clean
Acoustic environment	Sound booth
Evaluation procedure	Yes [6]

2.2 SIVA (ELRA)

The Italian speech corpus Speaker Identification and Verification Archives (SIVA) is composed of more than 2,000 calls collected over the public switched telephone network. The SIVA corpus consists of male and female users and male and female impostors. Speakers access the recording system by calling a toll free number. An automatic answering system guides them through the three sessions that make up a recording. In the first session, a list of 28 words (including digits and commands) is recorded using a standard enumerated prompt. The second session is a simple unidirectional dialog (the caller answers prompted questions) where personal information is asked (name, age, etc.). In the third session, the speaker is asked to read a continuous passage of phonetically balanced text that resembles a short curriculum vitae.

Table 2 SIVA Corpus Description

# of speakers	671 (335 M/336 F)
# sessions/speaker	1 - 26
Intersession interval	Days-months
Type of speech	Prompted words and digits, short questions and read text
Microphones	Variable telephone handsets
Channels	PSTN
Acoustic environment	Home/office
Evaluation procedure	Defined customers and imposters

2.3 PolyVar (ELRA)

PolyVar is a speaker verification corpus comprised of native and non-native speakers of French, mainly from Switzerland. It consists of read and spontaneous speech in Swiss and French amounting to 160 hours of speech. Thirty-one speakers called from 2 to 10 times and 41 speakers made more than 10 calls.

Table 3 PolyVar Corpus description

# of speakers	143 (85 M/58 F)
# sessions/speaker	1-229 (3600 total)
Intersession interval	Days-months
Type of speech	Read and prompted digits, word and sentences, questions, and spontaneous speech
Microphones	Variable telephone handsets
Channels	PSTN (Possibly ISDN)
Acoustic environment	Home/office
Evaluation procedure	No

2.4 POLYCOST (ELRA)

The POLYCOST corpus was collected under the COST 250 European project [7] for speaker verification. Most of the speech is non-native English with some speech in speaker’s native tongue covering 13 European countries. The speech was collected digitally over international ISDN telephone lines. The different languages in this corpus allow for experimentation on the effect of language on speaker recognition performance.

Table 4 POLYCOST Corpus Description

# of speakers	133 (74 M/59 F)
# sessions/speaker	>5
Intersession interval	Days-weeks
Type of speech	Fixed and prompted digit strings, read sentences, free monologue
Microphones	Variable telephone handsets
Channels	Digital ISDN
Acoustic environment	Home/office
Evaluation procedure	Yes [8] [9]

2.5 KING (LDC)

The KING corpus was collected at ITT in 1987 under a US Government research contract. The version now available from LDC, referred to as KING-92, is based on a 1992 reprocessing of the original recordings. It contains recorded speech from 51 male speakers in two versions, which differ in channel characteristics: one from a telephone handset and one from a high-quality microphone. The speakers are further subdivided into two groups, 25 in one and 26 in the other, who were recorded at different locations. For each speaker and channel, there are 10 files, corresponding to sessions of about 30 to 60 seconds duration each.

Table 5 KING Corpus Description

# of speakers	51 (all male)
# sessions/speaker	10
Intersession interval	Week-month
Type of speech	Extemporaneous descriptions of photograph to interlocutor
Microphones	Dual: Wideband and telephone handset (electret)
Channels	Dual: clean and PSTN
Acoustic environment	Sound booth
Evaluation procedure	Yes [10]

2.6 YOHO (LDC)

The YOHO corpus is designed to support text-dependent speaker verification evaluation for Government secure access applications. A high-quality telephone handset (Shure XTH-383) was used to collect the speech; however, the speech was not passed through a telephone channel. YOHO was recorded in a fairly quiet office environment with low-level office noise, fan noise, and occasional pages over a public address system. The LDC release of YOHO was designed to answer the question: does a speaker verification system perform at 0.1% false rejection rate and 0.01% false acceptance rate at 75% confidence with a 50% probability of passing the test? The phrases are randomized and prompted in a text-dependent speaker verification scenario using a “combination lock” phrase syntax. For example, a prompt could read: “Say: twenty-six, eighty-one, fifty-seven.” The Government withheld a portion of the corpus to validate performance claims [11].

Table 6 YOHO Corpus Description

# of speakers	138 (106 M/32 F)
# sessions/speaker	4 enrollments, 10 verifications
Intersession interval	Days-month (3 days nominal)
Type of speech	Prompted digit phrases
Microphones	Fixed high-quality in handset
Channels	3.8KHz/clean
Acoustic environment	Office
Evaluation procedure	Yes [11]

2.7 Switchboard I-II Including NIST Evaluation Subsets (LDC)

The Switchboard corpora, encompassing all phases, represents one of the largest collections of conversational, telephone speech recordings available. There are two main Switchboard corpora (I and II), two phases of Switchboard-II and several subsets of Switchboard I-II used to create the NIST speaker recognition evaluation corpora.

Both Switchboard-I and II were collected by a participant calling into an automated operator that connected him/her to another participant and recorded their conversation (as separate sides) for 5 minutes. The automated operator handled the information gathering and prompted callers with a topic to discuss. The main difference between Switchboard-I and II is the demographics of participants. In Switchboard-I, participants had a wide age and location distribution. In Switchboard-II, the participants were obtained from specific college campuses in different parts of the US for each phase. Next we briefly describe some of the aspects of the different speaker recognition corpora derived from Switchboard.

Table 7 Switchboard I-II Corpus Description

# of speakers	543 & 657 (~50% M/50% F)
# sessions/speaker	1-25 (5 min conversations)
Intersession interval	Days-weeks
Type of speech	Conversational
Microphones	Variable telephone handsets
Channels	PSTN
Acoustic environment	Home/Office
Evaluation procedure	Yes for NIST Eval sets [2]

SPIDRE: This corpus is a 2-CD subset of the Switchboard-I collection selected for speaker ID research, and with special attention to telephone instrument variation. It contains training and testing data for experiments in closed- or open-set identification or verification [12]. Combining the two sides of the conversations also permits speaker change detection or speaker monitoring experiments.

There are 45 “target” speakers; four conversations from each target are included, of which two are from the same handset. There are also 100 calls in which no target appears. Since all conversations are two-sided, this results in 180 target sides and 180 + 200 = 380 nontarget sides.

NIST Evaluation Corpus 1996: This corpus was derived from the entire Switchboard-I corpus for speaker verification evaluations [2]. The development data consisted of training and test speech from 45 male and 45 female speakers. The evaluation data consisted of training data from 21 male and 19 female target speakers plus test speech from these targets and

167 male and 216 female unseen imposters. The test data covered three durations (3, 10 and 30 sec) and was designed to be from matched and mismatched telephone handsets used during training by the targets [13].

NIST Evaluation Corpus 1997: This corpus was derived from the Switchboard-II Phase 1 corpus for speaker verification evaluations [2]. Development data for this evaluation was composed of the Eval96 development and evaluation data. This evaluation consisted of 400 targets (200 male/200 female) and 5000 test files for each of three durations (3, 10, 30 sec). Train and test data were also selected to allow examination of matched and mismatched telephone handset conditions [14].

NIST Evaluation Corpus 1998: This corpus was derived from the Switchboard-II Phase 2 corpus for speaker verification evaluations [2]. Development data for this evaluation was composed of the Eval96 and Eval97 development and evaluation data. This evaluation consisted of 500 targets (250 male/250 female) and 5000 test files for each of three durations (3, 10, 30 sec). Train and test data were also selected to allow examination of matched and mismatched telephone handset conditions [15].

A **multispeaker** version of the Eval98 corpus was created recently that consists of test segments containing speech from both speakers in a conversation. This corpus is useful for speaker-change-detection and speaker-location research.

Cellular Switchboard: A Switchboard-style corpus using cellular telephones has been designed. Collection is planned to begin in 1999 and it should be available in 2000. This will expand Switchboard from wireline to cellular telephony research and evaluation.

2.8 Speaker Recognition Corpus (OGI)

The Center for Spoken Language Understanding is collecting a large speech database for speaker recognition research. The initial release of the corpus contains approximately 100 speakers (a future release may contain 600 speakers) calling from different telephone environments and at different times. Each of these speakers calls OGI's system 12 times over a 2-year period. Speakers were asked to call from quiet and noisy locations and use various types of phones, such as cordless, cellular, and payphones. Several different types of data were requested from each speaker to provide a corpus useful for vocabulary-dependent and vocabulary-independent speaker identification and verification systems [16].

Table 8 OGI Speaker Recognition Corpus Description

# of speakers	100 (47 M/53 F)
# sessions/speaker	~12
Intersession interval	Months-2 years
Type of speech	Prompted phrases, digits, prompted monologue
Microphones	Variable telephone handsets
Channels	PSTN
Acoustic environment	Home/Office
Evaluation procedure	Under development

3. SUMMARY

We reviewed publicly available corpora suitable for speaker recognition research and evaluation. The list is not exhaustive, since we may have missed some corpora, but is representative of what is available. We hope this cataloging helps to increase awareness of their availability, their use, and participation in standard evaluations (e.g., the NIST Evaluations) to allow continued progress in the field of automatic speaker recognition. The authors express their gratitude to Håkan Melin, Mauro Falcone, Dominique Genoud, Ron Cole, Mike Noel, and Sadaoki Furui for their generous assistance.

4. REFERENCES

[1] Doddington, G. "Speaker Recognition Evaluation and Methodology: An Overview and Perspective," *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, Avignon, France, April 20-23, 1998, p. 60-66.

[2] NIST Speaker Recognition Evaluation Plans. http://www.nist.gov/speech/test.htm

[3] European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/

[4] Linguistic Data Consortium. http://www.ldc.upenn.edu/

[5] Oregon Graduate Institute. http://cslu.cse.ogi.edu/

[6] Reynolds, D., et al. "The Effects of Telephone Transmission Degradations on Speaker Recognition Performance," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1995, p. 329-332.

[7] Petrovska, D., et al. "POLYCOST: A Telephone-Speech Database for Speaker Recognition," *RLA2C,* Avignon, France, April 20-23, 1998, p. 211-214. http://www.speech.kth.se/~melin/papers/rla2c_ply.ps

[8] Melin, H. and J. Lindberg. "Guidelines for experiments on the POLYCOST database," *Proc. COST250 Workshop on The Application of Speaker Recognition Technologies in Telephony*, Vigo, Spain, Nov 11-12, 1996, p. 59-69. http://circwww.epfl.ch/polycost/baseline.htm

[9] Nordström, T., H. Melin, and J. Lindberg. "A Comparative Study of Speaker Verification Systems using the Polycost Database," To appear in: *Proceedings of International Conference on Spoken Language Proc (ICSLP)*, Sydney, Australia, Nov. 30 - Dec. 4, 1998.

[10] Reynolds, D. "Experimental Evaluation of Features for Robust Speaker Identification," *IEEE Trans on Speech and Audio Processing*, vol 2, Oct 1994, p. 639-643.

[11] Campbell, J. "Testing with The YOHO CD-ROM Voice Verification Corpus," *ICASSP*. Detroit, May 1995, p. 341-344. http://www.biometrics.org/REPORTS/ICASSP95.html

[12] Reynolds, D. "The Effects of Handset Variability on Speaker Recognition Performance: Experiment on the Switchboard Corpus," *ICASSP*, May 1996, p. 113-116.

[13] Reynolds, D. "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Eurospeech*, September 1997, p. 963-966.

[14] Pryzybocki, M. and A. Martin. "NIST Speaker Recognition Evaluation - 1997," *RLA2C,* Avignon, France, April 1998, p. 120-123.

[15] Pryzybocki, M. and A. Martin. "NIST Speaker Recognition Evaluations," *Proc. Int Conference on Language Resources and Evaluation (LREC)*, Grenada, Spain, May 1998, p. 331-335.

[16] Cole, R., M. Noel, and V. Noel. "The CSLU Speaker Recognition Corpus," To appear: *ICSLP*, Sydney, Australia, Nov. 30, 1998.