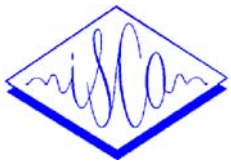
ISCA Archive
http://www.isca-speech.org/archive



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

THE CSLU SPEAKER RECOGNITION CORPUS

Ronald Cole, Mike Noel, Victoria Noel

Center for Spoken Language Understanding

Oregon Graduate Institute, USA

ABSTRACT

This paper describes the CSLU Speaker Recognition Corpus data collection. The corpus was motivated by a need for speech data from many speakers, under different environmental conditions, with each speaker providing data over a significant period of time. The corpus was designed to provide sufficient data to study phonetic variability within and across sessions, and to design and evaluate systems for both vocabulary independent and vocabulary specific recognition and verification tasks. The protocol includes fixed vocabulary phrases, digit strings, personal utterances (e.g., eye color), and fluent speech.

The resulting Speaker Recognition Corpus is a collection of telephone speech recordings from over 500 participants collected over a two-year period. We describe the data collection procedure, the protocol, the transcription methods and the current status of the Speaker Recognition Corpus.

1. OVERVIEW

Advances in speech technology require language resources to enable researchers to study and model the sources of variability in speech, and to develop, evaluate and compare systems across different sites. In the field of speaker recognition and verification, progress has been hampered by the lack of accessible language resources that reflect real-world conditions. A recent corpus developed by Campbell [1] provides a good starting point for speaker verification research, but it is limited in that the recordings were made in a quiet environment, and individual speakers were not recorded at different times over a period of several months. Furui has found that speaker verification performance degrades for a standard set of templates after only a few months [2]. In order to examine speaker changes over time in realistic environments, new corpora are needed.

Since 1996, CSLU has been developing a corpus for use in Speaker Recognition and Verification research. The corpus has been designed for use with both vocabulary dependent and vocabulary independent systems, and to support a variety of applications and research interests. The corpus is unique in that, for the first time, a significant amount of speech is collected and transcribed from a large set of speakers over a two-year period. Each speaker who participated in the data collection provided speech on 12 different occasions, from different telephones and locations, over a two year period. The callers provided a range of speech samples—personal information, phonetically rich phrases, digit strings and extemporaneous speech. In addition, some utterances were repeated 4 times, enabling analysis of within- and between session variability, as well the study of changes in speech over time. The final release of the corpus will have approximately 40 hours of recorded speech utterances from approximately 500 speakers. The corpus will contain word level transcriptions for most of the recorded utterances. Statistics concerning gender and age will accompany the corpus as well.

The data collection was initiated in September 1996. It is still underway, with the final set of speakers scheduled to complete their calls in July 1999. At present, 84 speakers have completed all twelve calls. The final release is scheduled during the fall of 1999. The initial release of the corpus, containing the completed, transcribed calls from 100 speakers, is scheduled for November 1998.

2. DATA COLLECTION

For this data collection, CSLU collected speech from each participant in twelve separate recording sessions over a two-year period. To normalize for seasonal effects such as colds in the winter or hayfever in the summer, we initiated the data collection for 12 different groups of subjects on successive months. The first group began in September 1996, the second group in October 1996, etc. The twelve calls made by each during the two year period occurred in 8 months, according to the following schedule: Year 1—month 1, two calls; month 4, one call; month 7, two calls; month 10, one call. The same schedule was then repeated for year 2.

To help assure continued participation and compliance with the goals of the data collection, CSLU staff sent letters and instruction packets to the participants prior to each call. This packet included specific instructions about the calls the participant was to make. In particular, during different sessions, participants were asked to call from quiet and noisy locations. They were also asked to use various types of phones such as cordless, cellular, and payphones.

Each participant called the data collection system twelve times in two-years. During each call they provided their name and an identification number. They were then prompted to answer questions, to repeat words and phrases, and to produce extemporaneous (“free”) speech on different topics. The following section describes the protocol in more detail.

3. PROTOCOL

One of the design goals for this corpus was to create a database that would be useful for both vocabulary dependent and vocabulary independent speaker recognition and verification systems. In order to meet this goal, several different types of data were requested from each speaker in the collection. In a couple of cases, the participants considered the information

requested too personal to be recorded. These participants were allowed to provide fictitious information, which they repeated each time they called.

3.1 Single Words

Participants were asked to repeat each of the following words four times during each call: *mango, choices, decision, whereabouts, azure, offstage.* These words were selected for their phonetic coverage. For example, *azure* contains the voiced postalveolar fricative which is rare in American English speech.

3.2 Phonetically Rich Phrases

Participants were asked to repeat each of the following phrases 4 times during each call.

1. *Joe books very few judges*
2. *It's been about two years since Davie kept shotguns*
3. *Tina got cued to make a quicker escape*
4. *Charlie did you think to measure the tree*
5. *Play in the street up ahead*
6. *Here I was in Miami and Illinois*
7. *Stop each car if it's little*
8. *A fifth wheel caught speeding*

The phrases were generated to provide phonetic combinations that are not frequent in American English. These phrases provide researchers with a set of utterances with good phonetic coverage that are each produced four times per session, for twelve sessions over a two year period.

3.3 Digit Strings

Many speaker recognition/verification systems require the user to produce a short, randomly generated digit string to read. The system uses the recording to make its decision about the user. To address these sorts of systems, the data collection system asked each participant to repeat the following digit strings.

5 3 8 2 4 *6 1 oh 9 7* *4 zero 7 1 3*

1 9 0 5 4 *2 8 3 7 6* *zero 5 2 3 9*

Most participants had no problem remembering the 5 digits produced during the prompt. However, in case of a mistake, the transcription reflects the actual digit string spoken rather than the expected string.

3.4 Personal Information

Many current systems query a user for personal information when performing speaker verification. To simulate this speakers were asked to say their mother's maiden name, their own name, their eye color, and the month in which they were born. In addition, during the first call for a speaker, the collection system asked them to invent a "personal password or passphrase". On subsequent calls, the speaker was asked to repeat their personal phrase.

3.5 Free Speech

Each speaker was asked to speak for about 20 seconds, describing one of his or her favorite items (e.g., favorite book, favorite movie, etc.). The "favorite" question was asked twice per call.

3.6 Mimic

The final prompt asked the caller to listen carefully as the prompter says the phrase: "If it doesn't matter who wins, why do we keep score?" The caller was asked to mimic that phrase and try to sound as much like the prompter as possible. This was only asked once per call.

4. PARTICIPANTS

4.1 Soliciting Participants

Our original goal was to collect 1200 speakers—12 groups of 100 speakers. To start the data collection, CSLU advertised for participants on the Internet. Within a few weeks we enlisted over 1500 volunteers. Since we expected some participants to drop out of the data collection over time, the additional 300 participants were added as "padding" to the twelve groups.

Once the data collection began, it became apparent that the dropout rates were much higher than expected for each group. In some cases, the drop out rates reached 87% (see table below). In order to maintain a large number of participants, who would complete two years of recording, CSLU solicited more participants and padded the groups that had not yet begun their sessions. In addition, to maintain participation, we developed an incentive program. Each subject was promised a $5 gift certificate upon completion of each call and a $20 gift certificate bonus at the halfway mark, and after all calls were made. Additionally, each participant was promised that if they completed all twelve calls, they would be entered into a drawing for $10,000 cash at the end of the data collection (summer 1999).

As of April 98, 676 participants were still active in the project. Experience indicates that these numbers will decrease over the next year but we do not expect them to fall below a total of 500 participants.

Group	Start #	# of drops	Current #	Drop Rate
A	112	52	60	46%
B	128	82	46	64%
C	112	79	33	70%
D	120	98	22	82%
E	115	85	30	74%
F	211	141	70	67%
G	219	145	74	66%
H	187	140	47	75%

I	225	143	82	64%
J	371	298	73	80%
K	363	314	49	87%
L	422	332	90	79%
Total	2585	1909	676	

Table 1: Attrition rate for each group, and number of current participants. The dropout rate is computed as the percentage of participants in each group who did not complete all of their calls.

4.2 Gender Balance

Originally, each group was formed with an equal number of male and female participants, however, as subjects were dropped from the project, the balance was not maintained. Currently, 47% of the remaining participants are male and 53% are female.

4.3 Age Distribution

As participants signed up for the project they provided information about their age group. The following table shows the distribution of age groups in the study.

Age	Percentage
10-15	3%
16-20	9%
21-	32%
31-	29%
41-	16%
51-	5%
61-	2%
unknown	4%

4.4 Geographic Distribution

The following table shows the distribution of participants throughout the United States.

Percentage	Cumulative	State Abbreviation
10%	10%	CA
6%	16%	NY
5%	21%	PA
4%	29%	FL TX
3%	50%	AZ IL MA OH OR VA WA
<3%	100%	AK AL AR CO CT DE GA IA ID IN KS KY LA MD ME MI MN MO MS MT NC ND NH NJ NM OK SC SD TN UT WV WY

Table 2: Percentage and Cumulative Percentage of participants from the listed states.

5. TRANSCRIPTION

Each utterance in the corpus is transcribed either by a human or by a machine. Normally, CSLU corpora are transcribed by our professional staff, but due to the immense amount of data and the desire to release the data as soon as possible, speech recognizers files were used to transcribe portions of the corpus. Files that were transcribed automatically were assigned a confidence score, and files that were flagged as low confidence were inspected by human transcribers.

5.1 Human Transcriptions

All manual transcriptions were produced according to the CSLU transcription conventions [3]. After transcription, the following quality control checks were performed.

Files were checked for proper convention usage. Each file was checked automatically for improper use of transcription conventions. For example, cut-off speech markers not connected to a word. Any malformed transcriptions were flagged and investigated by a transcriber.

Files were checked for spelling. A simple spell checker was run over all of the manually produced transcriptions. In addition to checking the spelling of normal words, consistency between all of a speaker's utterances was checked.

5.2 Machine Transcriptions

Because of the number of files needing transcriptions, CSLU decided to use a recognizer to automatically transcribe portions of the corpus. The recognizer was provided with the expected transcription and the speech recording. It then generated a confidence score that indicated how likely it was that the utterance matched the transcription. Any utterances with confidence below a threshold, selected from training data, were flagged for manual inspection and transcription. (In the release of the corpus, files transcribed manually vs. automatically are distinguished by file extension.)

6. TRAINING AND TEST SETS

CSLU envisions this database being used by researchers to study variability in speech and speakers within a single session, across sessions over a significant period of time, and as a function of different communication channels and environments. In addition, the corpus is designed to support research, development and evaluation of speech recognition and speaker recognition systems. In order to facilitate comparison of experimental results across sites, we have specified training, development test, and final test sets within the data.

7. DISTRIBUTION AND RELEASES

The first release of the Speaker Recognition Corpus, in November, 1998, will contain two year's worth of transcribed speech from over 100 speakers. Over the course of the next year, the speech from the remaining speakers will be packaged for incremental releases. The corpus is freely available from

CSLU to educational and non-profit organizations. All others should contact CSLU for distribution policy.

More information about CSLU corpora and distribution policies can be found at *http://www.cse.ogi.edu/cslu*.

8. REFERENCES

1. J. P. Campbell, Jr. Testing with the YOHO CD-ROM voice verification corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 341-344, Detroit, MI, 1995.

2. S. Furui, An Overview of speaker recognition technology. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pages 1-9, Martigny, Switzerland, April 1994.

3. T. Lander, *The CSLU Labeling Guide*, CSLU, Oregon, June 1996

9. ACKNOWLEDGEMENTS

This data collection was supported by a grant from NSF (NSF IRI-9529006 Human Language Resources/Multilanguage Systems) and CSLU center member support. The views expressed in this paper do not necessarily represent the views of NSF.

CSLU would also like to thank the callers who have participated in this data collection. It goes without saying that without their contribution, this corpus wouldn't be nearly so interesting.