ESCA Workshop on Automatic Speaker
Recognition, Identification, and Verification
Martigny, Switzerland
April 5-7, 1994

Public Databases for Speaker Recognition and Verification

John Godfrey, David Graff, Alvin Martin

Abstract—

In this paper we review several major speech corpora which are designed to support research in speaker recognition and related areas: the KING corpus; KING-SAM, a derivative of the KING corpus; the YOHO corpus; the SWITCHBOARD corpus; and SPIDRE, a derivative subset of SWITCHBOARD. Each one has design characteristics which make it more appropriate for certain types of research or technology development than others. Our purpose here is to acquaint researchers with these properties so that they can make the best choice for their purposes. We will attempt to highlight the amount and nature of the speech data in each corpus, its intended use for training or test where applicable, and the strengths and limitations of each dataset for research and development in such areas as speaker identification, speaker verification, and speaker monitoring.

Keywords— **Speaker verification, speaker authentication, speech databases**

1. Introduction

The importance of publicly accessible corpora for research and development is now widely recognized. In basic science it has long been a principle that experiments must be replicable for results to be accepted and pursued by other investigators. In technology development as well, where substantial investments of resources are often at risk, it is usually impossible to evaluate claims of performance unless they are tested on data that competitors can also use.

In this paper we review several major speech corpora which are designed to support research in speaker recognition, speaker verification, or related areas: the KING corpus, now being published in a revised version known as KING-92; KING-SAM, a derivative of the KING corpus; the YOHO corpus; the SWITCHBOARD corpus; and SPIDRE, a derivative subset of SWITCHBOARD. Each one has design characteristics which make it more appropriate for certain types of research or technology development than others, and all of them are, or shortly will be, available publicly from the Linguistic Data Consortium (LDC). We will attempt to highlight the strengths and limitations of each dataset, as well as their intended subdivisions into training and test, development and evaluation.

2. Database Descriptions

A summary description of the five databases is provided below, indicating collection protocols, speaker population, audio quality, etc., and their intended applications.

2.1 KING-92

The KING corpus was collected at ITT in 1987 by Alan Higgins under a US government research contract, and although other contractors have received it, it has not been officially available for public use before now. 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 ten files, corresponding to sessions of about 30 to 60 seconds' duration each. The interval between sessions varies from a week to a month. KING is designed principally for closed set experiments in text-independent speaker identification or verification over "toll-quality" telephone lines, although the single-sided collection format does not permit simulation of real telephone traffic. The ten sessions allow for a variety of divisions into training and test data, with the possibility of multiple test sets. For example, one could examine the effects of the amount of training on performance, or examine the variability of performance over several test samples (sessions) given a fixed amount of training (but see below about the "Great Divide".)

The collection method used in KING was to establish a call from a laboratory location at ITT (either San Diego, CA or Nutley, NJ) over long distance lines and back to another phone at the same location. The phones used by the test subjects were equipped with an additional microphone, so two parallel recordings were made of that side of the conversation, while the interlocutor's side was not recorded. The two parties either spoke spontaneously or carried out a variety of tasks designed to elicit natural-sounding speech: interpreting a drawing, solving a problem, describing a picture, etc.

There were 25 speakers in Nutley and 26 in San Diego. Speech-to-noise ratios average about 10 dB worse for the Nutley telephone data than for San Diego; in fact it is less than 20 dB for over half the Nutley files. Users of this corpus therefore usually run separate experiments, or at least report results separately, according to site. A more subtle difference in the recordings, however, sometimes referred to as the "Great Divide," cuts across the telephone data for the San Diego speakers. This was apparently due to a

J. Godfrey is Executive Director of the Linguistic Data Consortium (LDC) at the University of Pennsylvania in Philadelphia, PA. E-mail: jgodfrey@unagi.cis.upenn.edu

D. Graff is Programmer Analyst for the LDC. E-mail: graff@chestnut.ling.upenn.edu

A. Martin is with the National Institute of Standards and Technology (NIST) in Gaithersburg, MD. E-mail: alvin@jaguar.ncsl.nist.gov

minor equipment change which was made during the collection; it results in a slight but consistent change in the average long term spectrum of the telephone data recorded after the fifth session. Training and testing on data from the same side of this divide gives significantly better results than across it. Since the discovery of this difference, investigators now generally report results on the first and last five sessions of the San Diego telephone KING data separately, or they report within vs. across this boundary. A detailed description of the spectral differences will be found in a report by Thomas Crystal and Ned Neuburg which accompanies the CD-ROM version.

Since there are a number of published papers with results based on the original KING corpus [8], [10], [5], [2], [7], [6], and two versions of the data in existence, it is worth noting here that the new CD-ROM version, called KING-92, is based on a 1992 re-issue of the data from ITT. It differs from the original corpus in a few details, probably not enough to affect results significantly:

- The original data was sampled at 10 kHz, but has now been resampled at 8 kHz;
- Missing segments, most on the order of seconds, have been restored to the data, and the alignment between the high quality microphone and the telephone handset data files has been corrected;
- Originally both an orthographic and a phonetic transcription of the data, with time alignments, were part of the corpus, but there were numerous errors; only an unaligned orthographic transcription has been retained.
- Documentation has been changed to reflect these differences, and a description of the artifactual division between sessions 1-5 and 6-10 in the San Diego telephone data is included.

At this writing, LDC plans to publish the CD-ROM version of KING-92 as a two-disk set, one containing the wideband data and the other the narrowband, by April 1994.

2.2 KING-SAM

A derivative of KING is the "KING-SAM" (for Speaker Authentication Monitoring) corpus, created by the speech group at the National Institute of Standards and Technology (NIST) in 1993. KING-SAM uses speech data from the KING Corpus, and is intended to support research on automatic detection of speaker change, or "speaker monitoring." A typical task is to sample continuously and automatically report whether the known or authorized voice on a channel has changed. This would be a natural extension, for example, of technology which uses voice verification to authorize financial transactions by phone.

Each CD consists of spliced segments containing parts of two sessions (which are 30 to 60 seconds each) from the KING Corpus. Half the segments involve different sessions of the same speaker, and half involve sessions of different speakers. In addition, two whole KING sessions for each speaker (102 total sessions) are included as data that may be used to make speaker models. Each other session, with an appropriate division point chosen, was used twice: once in initial and final segments involving different speakers, and once in initial and final segments involving different sessions of the same speaker. In general, the four sessions involved in pairings with a given session are distinct. This results in 796 paired segments averaging about 40 seconds each. With the model-building sessions, this makes about ten hours of data, which at 8 KHz with 16-bit samples nearly fills a CD.

Each spliced segment contains at least ten seconds of speech from each of its component sessions, so the first ten seconds may safely be used to create a model of the initial speaker. The speakers are either both New Jersey speakers or both San Diego speakers. The splice point is always at the end of a short silence. The segment from the first speaker concludes with silence; the segment from the second speaker begins with speech. The speech gains of the segments of each spliced cut are equalized, and identical speech spectrum shaped white noise is added to each part to equalize the snrs of the parts to within 3 dB, while maintaining an overall snr for each segment of at least 20 dB. (usually considerably higher). All of this is designed to make the splice points less apparent.

Splice points were chosen at the end of the longest silence in a session for which there is at least ten seconds of signal on each side. This makes the distribution of segment lengths, as fractions of the total durations between ten seconds after the start and ten seconds before the end of the sessions, roughly uniform. A chi-square (95% significance) test has been applied to verify this uniformity.

The data used is the high-quality microphone speech version of KING-92, filtered to telephone bandwidth. This version of the speech data was used to avoid telephone channel differences which might cue a monitoring algorithm. This also avoids the channel anomaly between the first and last five San Diego sessions in the telephone data.

The KING-SAM Corpus is divided into two CDs which are similarly designed so that they may either be combined or used separately as a development and an evaluation corpus. For each CD, half the data has been designated to be for training and development, and half for test, for users wanting such a division. The names of the spliced segments do not indicate the speaker(s) used to create them. Rather, for each half of each CD there is a "key", specifying speakers and splice points for each spliced segment. At this writing, LDC plans to release the KING-SAM corpus by April 1994.

2.3 YOHO

The YOHO corpus, like KING, was collected at ITT in 1989 by Alan Higgins under a US government research contract [4], [1]. Its purpose is to support text-dependent speaker authentication research such as is used in "secure access" technology. There are 186 speakers (156 male, 30 female), each of whom completed between 4 and 13 recording sessions. In each session, the speaker was prompted with a series of "combination-lock" phrases to be read aloud; each phrase was a sequence of three two-digit numbers (e.g. "35 - 72 - 41", pronounced "thirty-five seventytwo forty-one"). The first 4 sessions for a given speaker were enrollment sessions of 24 phrases, and all additional sessions were verification trials of 4 phrases each. The total collection contains 744 enrollment sessions, and 1919 trial sessions. The nominal time interval between sessions for a given speaker was three days.

The "secure access" scenario, though commonly involving a telephone handset, need not involve local telephone circuits. In YOHO, for example, all sessions took place over the same handset, and the recording setup did not use commercial telephone lines. Moreover, although sampling was done at 8 kHz, with 12 bit resolution, the signal was bandpass filtered, with a high frequency cutoff of 3800 Hz and a low frequency cutoff of about 120 Hz; thus its bandwidth is wider than telephone speech data.

YOHO should be available from LDC in 2 CD-ROMs by May 1994.

2.4 SWITCHBOARD

SWITCHBOARD is a large multispeaker corpus of telephone conversations. Although designed to support several types of speech and language research, its variety of speakers, speech data, telephone handsets, and recording conditions make SWITCHBOARD a rich source for speaker verification experiments of several kinds. [5], [9], [12]

Collected at Texas Instruments with funding from ARPA, SWITCHBOARD includes about 2430 conversations averaging 6 minutes in length; in other terms, over 240 hours of recorded and transcribed speech, about 3 million words, spoken by over 500 speakers of both sexes from every major dialect of American English. [3] The data is 8 kHz, 8-bit mu-law encoded, with the two channels interleaved in each audio file.

In addition to its volume, SWITCHBOARD has a number of unique features contributing to its value for telephone-based speaker identification technology development.

- SWITCHBOARD was collected without human intervention, under computer control. From a human factors perspective, automation guards against the intrusion of experimenter bias, and guarantees a degree of uniformity throughout the long period of data collection. The protocols were further intended to elicit natural and spontaneous speech by the participants. The transcribers' ratings indicate that they perceived the conversations as highly natural.
- The use of T1 lines and automatic switching software made it possible to collect the digital version of the speech signals directly from the telephone network, and also to isolate the two sides of the conversations. The goal was to have real telephone speech, routed through the public network, but with no degradation due to the collection system. Isolation of the callers, within the limits of network echo cancelling performance, permits training on each speaker's voice separately, and then testing on either one or both speakers in any conversation.
- The speech is fully transcribed, and the transcription conventions documented.
- Each transcript is accompanied by a time alignment file, which estimates the beginning time and duration of each word in the transcript in centiseconds. The time alignment was accomplished with supervised phone-based speech recognition, as described by Wheatley et al. [11] The corpus is therefore capable of supporting not only purely text-independent approaches to speaker verification, but also those which make use of any degree of knowledge of the text, including phonetics.
- SWITCHBOARD has both depth and breadth of coverage for studying speaker characteristics. Forty eight people participated 20 times or more; this yields at least an hour of speech, enough for extensive training or modeling and for repeated testing with unseen material. Hundreds of others participated ten times or less, providing a pool large enough for many open-set experiments.
- The participants' demographics, as well as the dates, times, and other pertinent information about each phone call, are recorded in relational database tables. Except for personal information about the callers, these tables are included with the corpus. The volunteers who participated provided information relevant to studies of voice, dialect, and other aspects of speech style, including age, sex, education, current residence, and places of residence during formative years. The exact time and the area code of origin of each call is provided, as well as a means of telling which calls by the same person came from different telephones.

SWITCHBOARD is available on 26 CD-ROMs (25 of speech data, 1 of text and alignments) from LDC.

2.5 SPIDRE

For evaluation purposes, SWITCHBOARD has certain limitations as a speaker verification corpus. The most obvious is the amount of data (240 hours of speech, over 12 Gigabytes), which is an obstacle to many researchers. There are also so many ways of using less than the entire corpus that no two investigators are likely to do similar experiments independently. Also, although many callers participated in enough calls to be used as verification targets, they typically did not use more than two or three different telephones. Since the signal quality, due to the all-digital collection scheme, is very high in SWITCHBOARD, the handset assumes even more importance as a determinant of the channel characteristics. But the telephone number used by each party in each call is the only evidence we have of handset identity, so the only thing that can be said is whether two samples are from the same or different instruments, not whether they contained the same type of microphone. Thus this important dimension of variability cannot easily be studied.

For these and other reasons it was decided to create at least one derivative subcorpus of manageable size, with a structure specifically conducive to SPeaker IDentification REsearch (SPIDRE), and with special attention to telephone instrument variation. Within limits imposed by its smaller size (2 CD-ROMs, about 1.2 Gbytes of data), SPIDRE provides training and test data for experiments on both closed and open set speaker recognition, speaker verification, and even speaker monitoring (or speaker change detection).

The selection of data for SPIDRE from the SWITCHBOARD corpus was done as follows:

- callers who used at least three different telephones (= 45 targets)
- one call each from two telephones (= 90 calls)
- two calls from a third telephone (= 90 calls)
- both sides of the calls (= 180 nontargets)
- another 100 calls by other callers (= 200 nontargets)

In other words, there are four sets of conversations from 45 target speakers; sets one and two are "same phone" data for each speaker; sets three and four have data from "different phones", and are different from each other. In each conversation, the target speaker is paired with a different nontarget. Transcripts and alignment files permit automatic location of the speech by each talker, including places where both speak simultaneously.

An example of the possible uses of this corpus would be: train on one conversation per target speaker; test on another conversation by that speaker on the same telephone; test again on another conversation by that speaker on a different telephone, and again on the data from the third telephone. In preliminary tests of a closed set experiment like this, an algorithm which produced an error rate of less than 7% on the "same phone" data produced error rates of 40% and 49% on "different phone" data.

SPIDRE is available from LDC as a two-disk set, in which each CD contains half of both the target and nontarget data, so that they can be used separately as development and evaluation corpora or combined to create one larger corpus.

3. Conclusions

We conclude with a summary of the best uses of each of the corpora described above.

KING-92 best supports closed set speaker recognition by conventional spectrum-based methods. It permits repeated experiments over two datasets with different average channel qualities, probably typical of analog telephone circuits in the US. It has enough data for several develpment/test cycles in the Nutley speaker set, and two subsets with different channel characteristics in the San Diego set. The use of all males of similar ages and demographics makes the voices more challenging. The one-sided recording, the varied styles of speech from session to session, and the uncertain quality of the transcripts makes KING inappropriate for speech recognition based techniques.

KING-SAM is an artificial creation, designed to permit speaker monitoring research to begin on a manageable data set with well known characteristics, and structured for easy scoring and evaluation.

YOHO supports only fixed-phrase speaker authentication studies. The large number of speakers and the systematic set of impostor utterances makes it ideal for this type of study. Although telephone-like in bandwidth, its lack of handset or channel variation must be kept in mind if technology development is the research goal.

SWITCHBOARD, because of its large size and underlying database of information about calls and callers, permits researchers to design many types of experiments, either closed or open set. The fact that the two sides of a call are available separately or combined makes speaker change detection (monitoring) studies feasible. The large number of voices and high quality signal invites in-depth studies of voice characteristics. The time-aligned transcripts can support unorthodox approaches, such as speech recognition-based techniques. The principal disadvantages for telephone-based applications are the limited variation of handsets per speaker and possible ceiling effects due to the atypically high channel quality.

SPIDRE attempts to remedy the ceiling effects and logistical problems of SWITCHBOARD by selecting a subset in which handset variation is a main variable. Handset differences are not documented, however; only the phone number is known. SPIDRE is large enough to support two cycles (development and evaluation) of testing for most types of speaker verification experiments, as well as speaker change detection, on a smaller scale than SWITCHBOARD but not much smaller than KING (about 10 targets per sex on each of two CDs).

References

[1] J. P. Campbell, Jr., *Features and Measures for Speaker Recognition*, Ph.D. dissertation, Oklahoma State University, 1992.

[2] H.Gish, "Robust Discrimination in Automatic Speaker Identification," ICASSP-90 pp. 289-292

[3] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," Proc. ICASSP-92, Vol I, 517-520, 1992.

[4] A. Higgins, et al., *YOHO Speaker Authentication Final Report*, ITT Defense Communications Division, 1989.

[5] A. Higgins, et. al, "Voice Identification Using Nearest Neighbor Distance Measure," ICASSP-93 pp. II-375-II-378

[6] Y. Kao, et. al, "Free-Text Speaker Identification Over Long Distance Telephone Channel Using Hypothesized Phonetic Segmentation," ICASSP-92 pp. II-177 - II-180

[7] Y. Kao, et. al, "Robustness Study of Free-Text Speaker Identification and Verification," ICASSP-93 pp. II-379–II-382

[8] D.A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, PhD Thesis, Georgia Institute of Technology, August 1992

[9] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 5-7 1994

[10] D.A. Reynolds and R.C. Rose, "An Integrated Speech-Background Model for Robust Speaker Identification," ICASSP-92 pp. II-185 - II-188

[11] B. Wheatley, G. Doddington, C. Hemphill, J. Godfrey, E.C. Holliman, J. McDaniel, and D. Fisher, "Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech," Proc. ICASSP-92, pp. I-533 - I-536, 1992.

[12] G. Yu and H. Gish, "Identification of Speakers Engaged in Dialog," ICASSP-93 pp. II-383 - II-386