THE COLLECTION OF TWO SPEAKER RECOGNITION TARGETED SPEECH DATABASES

Michael Barlow, Ian Booth, and Andrew Parr

Speaker Verification Group,
Department of Electrical Engineering
University of Queensland

ABSTRACT - In recent years speech databases such as TIMIT have become available to the general research community. Such databases as are currently available are designed specifically with automatic speech recognition research in mind and as such are deficient in a number of aspects for automatic speaker recognition research; chiefly capturing repeated utterances from a large number of speakers over the long-term. The design and acquisition of two speech databases specifically for speaker recognition research are described, including an access system to a microcomputer laboratory.

INTRODUCTION

Speech research is data driven. A suitably large and spanning corpus of utterances is required to design, test and evaluate automatic approaches to such problems as speech and speaker recognition. Lacking adequate speech material, inconclusive or even artificially inflated results, due to lack of generality, will often be derived. Therefore, a necessary task when conducting significant speech research is the acquisition of a speech database suitable for the task being investigated.

In recent years a number of large databases, the best known commercial example being TIMIT (Lamel, Kassel & Seneff, 1986), have become available outside of the laboratories in which they originated. Such databases have eased the task for speech recognition researchers; as the time consuming and expensive task of data collection is often no longer necessary; as well as providing a common frame of reference for comparing and contrasting different approaches to the same problem.

Unfortunately these databases have been designed primarily with the speech recognition task in mind and are poorly suited for research into speaker recognition or speaker characteristics in general. The primary characteristics of a speech database for speaker recognition or speaker characteristic investigation are:- (1) A large number of speakers be represented; (2) A number of recordings of each speaker be made to capture intra-speaker variability; (3) Common and suitable utterances be recorded by all speakers; (4) The recorded speaker population adequately models that of the general population or the particular subgroup of the speaker population being examined.

In particular, capturing multiple recordings from each speaker over a period of weeks and months is a need for speaker recognition databases which currently available databases targeted for speech recognition do not address. It is a well know phenomenon that the acoustic realisation of an utterance by a given speaker changes from repetition to repetition. Further, researchers such as Furui et. al. (1972) have shown that adult speakers' voices show non-trivial changes after periods of 3 months or more. These two sources of variance:- long-term and short-term ensure that multiple samples of a speaker's voices over a period of several months are necessary to adequately capture the true variance of the speaker's voice.

BACKGROUND

In 1991 a three year externally funded project to implement a commercial speaker verification system was undertaken in the department of Electrical Engineering at the University of Queensland. At the inception of the project no speech database then available was found to meet the requirements of an examination of the speaker verification problem. It was therefore necessary to collect a database of speakers that would allow the evaluation of the automatic speaker recognition systems being designed.

Critical properties pertaining to any speech data collected were felt to be:-

- A large number of speakers be collected.
- Speakers should be sampled over many months.
- Different utterances (phonetic sequences) be obtained.
- Specific sources of intra-speaker variability be targeted.
- Speech samples be obtained under quiet and noisy ("real-word" conditions).
- Speakers sampled should be available for the duration of the project and taken as a group show distribution properties corresponding to the general population.

Several of these properties were found to be complementary or mutually exclusive. It was therefore decided to collect two databases, the larger one representing speech from several hundred individuals under "real-world" conditions, and the second from a smaller set of 40 speakers under quiet conditions for whom specific sources of intra-speaker variability could be explored.

LABORATORY ACCESS DATABASE

In order to acquire data and test speaker verification systems under real application conditions, the primary database collected was implemented as part of an access system to a laboratory of personal computers used by several hundred undergraduate students in the department of electrical engineering. Primary goals in the collection of this database have been the collection of speech samples from a large number of speakers under real application conditions.

The door monitoring system further provides secure access control for the PC laboratory; logging and authorising entries and exits; an important role given the equipment present in the laboratory (over 50 PCs) and the number of students using the laboratory daily (several hundred during semester).

System Configuration

At the core of the door monitoring system for the acquisition of speech samples is a commercial card-swipe access control system; providing low level functions such as card reading, pin entry, and door lock/unlock facilities. In-house developed software and hardware, sitting atop the commercial system, provides the higher level functions of speech acquisition, database maintenance and speaker verification. Figure 1 provides a schematic of the system configuration.

A PC driven system monitors the commercial access control system. When a valid entry attempt is detected (card swipe) the PC audibly prompts the student for input via a wall mounted speaker. Speaker utterances are captured via a wall mounted microphone, digitised, and sent over the ethernet to the main processing facility; a Silicon Graphics 4D/440. Computational or storage intensive tasks such as end-point detection, database management and speaker verification are carried out on the 4D/440 (the server) and the results communicated back to the PC (the client).

System design allows the commercial door monitoring system to act in a stand-alone (pass through) mode. If software errors such as database inconsistencies or timeouts over the ethernet are detected the client code resigns control of the door to the commercial system. Power failures for the PC are handled in a similar manner.

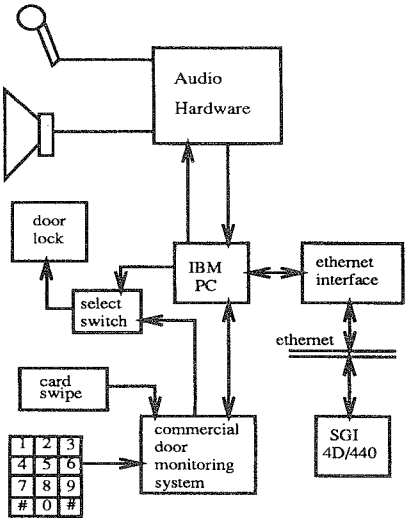

Figure 1: Schematic diagram of door-lock access control system for the acquisition of speech data. A Commercial door monitoring system is interfaced to a PC which communicates via ethernet with a larger machine responsible for database maintenance.

A hierarchical database management system maintains information on students, entries and exits from the laboratory, as well as all speech samples collected via the system. Simple queries of the database provide powerful supervisor capabilities as well as a means of accessing the stored speech data.

Speech Data

Speech data is acquired by the door monitoring system using a wall mounted AKG-D321 microphone. Utterances are sampled at 16kHz, low-pass filtered at 7.8 kHz via a 12'th order Butterworth filter before being quantised at 16 bits.

Over three hundred (300) users (students) are registered with the system. Upon each valid entry attempt students are prompted for a number of isolated utterances. Utterance selection for the prompting process is a simple rule driven procedure allowing easy variance of the number and type of utterances prompted for. The criteria for utterance selection may be altered without halting system operation.

Currently the system is configured to prompt for two random digits upon each entry by a student. With an average number of 70 students per day using the laboratory during semester; an average of 700 speech samples per week are collected.

Preliminary analysis of the average noise levels for recordings show a range of SNR values from 2 to 4. Speech babble, room reverberation, and impulsive noise sources (such as door slams) are the primary maskers of the speech signal.

DEPARTMENTAL DATABASE

To complement the large, application-like database collected via the laboratory access system, a second smaller database sampled under quiet conditions, with greater linguistic/phonemic variability is being collected. The primary issues addressed in the design of this database are sources of intra-speaker variability

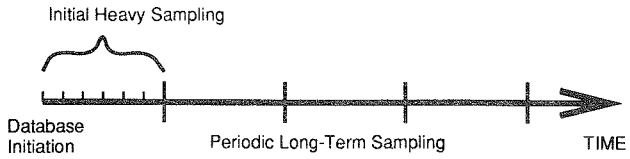

Figure 2: Sampling frequency strategy for the departmental database. Sampling is initially consistent to capture a representative pool of samples from each speaker. Sampling frequency is then decreased, though sampling is continued so long-term variability may be examined.

and choice of a good utterance set for speaker recognition. Further, the smaller database allows new approaches and modifications to be tested rapidly before being more fully explored with the large laboratory access database.

Speaker Population

Due to the need for a high degree of co-operation from participating speakers, members of the Department of Electrical Engineering were invited to participate in the collection of data. Such a speaker population is highly stable, easily contacted at short notice, diverse, and quite co-operative.

A total of 42 departmental members comprise the speaker population. The population is comprised of 6 females and 38 males ranging in age from 20 to 61 years with a mean age of 30.6. Population members show a high degree of educational, vocational and cultural diversity. Prior to recordings all speakers completed an information sheet detailing factors deemed significant to their language development. The collected information is highly analogous to that specified in the National Spoken Language Database (NSLD) project (Millar, Dermody, Harrington, & Vonwiller, 1990).

Recording Strategy

Members of the speaker population are recorded individually in a quiet office used solely for that purpose. Speakers wear a close-talking (headset) Sennheiser HMD 410 microphone and are recorded directly to a Sony TCD-D10 portable Digital-Audio-Tape machine. Recordings are then transferred to computer and stored on disk with a down-sampled rate of 16kHz and 16 bit quantisation.

For each recording session speakers are provided with a set of visual prompts specifying the utterances to record. A supervisor is present at each recording to ensure the speaker successfully completes the recording task and to provide assistance as required.

All speakers record a particular set of utterances within one week of each other. However, in addition to the basic recordings a subset of 11 young male speakers have been designated as a high repetition group. These speakers have made an additional number of recordings of the material and are also intended as the test population for induced sources of intra-speaker variability such as exertion.

The basic sampling strategy concerning the spacing of recording sessions is shown in Figure 2. As can be seen initial sampling seeks to capture a minimum representative pool of samples for each speaker. Sampling frequency is then decreased in order that long-term variability be examined without unduly inconveniencing speakers.

Speech Data

Speakers are asked to record a number of different utterance sets ranging from isolated words through to read texts and spontaneous speech. The variety of utterances recorded allows a range of possible techniques and applications of speaker recognition be explored.

At the time of paper authorship the speaker population has recorded or is in the process of recording the following different utterance sets:-

- Isolated digits
- PB (Phonetically Balanced) word list
- PB sentence set
- A read passage
- Extemporaneous speech.

It is intended that items such as the read passage (Rainbow passage) or the extemporaneous speech (speakers asked to describe how they arrived at work that day) be recorded once only. Such passages serve to more fully capture a speaker's range of utterances than isolated words or sentence lists alone. The word and sentence lists are repeated at each recording session, providing a large database upon which to perform application targeted experimentation.

CONCLUSION

A number of commercial speech databases are now becoming available. Unfortunately these databases are chiefly designed for speech recognition research and do not adequately meet the needs of speaker recognition work. In particular such databases as are currently available lack a large number of speakers, and a number of repeated samples over a period of months from each speaker. To meet the needs for the design of a speaker verification system the Speaker Verification Group at the University of Queensland began collecting speech data suitable for speaker recognition research.

It was decided that speech data collected must meet certain properties:- (1) A large number of speakers; (2) Multiple repetitions of material from all speakers; (3) Speakers sampled over a period of months; (4) Data be acquired under quiet and real-world noise level conditions; (5) Specific instance of intra-speaker variability be targeted; (6) Different phonemic material be represented; (7) The speaker population model potential populations for a commercial application of speaker verification. To meet these diverse needs two databases are being collected.

The first and largest database is being acquired through the mechanism of a laboratory access system. A laboratory of microcomputers is used by several hundred undergraduate students in the department of electrical engineering at the University of Queensland. A combined hardware-software system controls access to the laboratory; requiring students to provide a number of speech samples on each entry to the laboratory. Via this mechanism a large number of speech samples from a speaker population in excess of 300 is being acquired under real-world application conditions.

The second and smaller database of over 40 speakers is designed to complement the larger database. Speakers record a larger variety of material ranging from isolated words through to read texts and extemporaneous speech. Speakers are recorded under quiet conditions with a close-talking microphone and digital audio tape.

Combined, the two databases address the specification requirements of the speaker verification group. The large laboratory access database allows the exploration of noise reduction issues and system performance for large speaker population sizes while the smaller database allows the examination of sources of intra-speaker variability and utterance selection.

Both databases are maintained in a format closely consistent with the specifications of the National Spoken Language Database (NSLD). It is the intention of the group to ultimately make the data available to other researchers.

REFERENCES

Furui, S., Itakura, F., & Saito, S. (1972) *Talker Recognition by Longtime Averaged Speech Spectra*, Electronics and Communication in Japan, 55-A(10), 54-61.

Lamel, L.F., Kassel, R.H. & Seneff, S. (1986) *Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus*, Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, 100-109.

Millar, J.B., Dermody, P., Harrington, J.M., & Vonwiller, J. (1990) *A National Cluster of Spoken Language Databases for Australia*, Proc. Third Aust. Int. Conf. Speech Science and Technology, 440-445.