BELL LABS CONNECTED DIGIT DATABASES FOR TELEPHONE SPEECH RECOGNITION

Qiru Zhou, Imed Zitouni, Qi Li[1]

Bell Labs, Lucent Technologies
600 Mountain Ave., Murray Hill, NJ 07974, USA
{qzhou, zitouni}@research.bell-labs.com
[1]Li Creative Technologies, Inc. New Providence, NJ, USA
qili@ieee.org

ABSTRACT

This paper describes Bell Labs Connected Digits databases (BLCD), which were collected over the landline telephone networks. The BLCD databases were designed to provide a standard benchmark for evaluating the performances of different connected digit recognition systems. It is also a vehicle for research and diagnosis of specific problems in automatic connected digit recognition. We first describe the content and the organization of the BLCD databases, and then present an automatic database verification procedure utilizing automatic speech recognition (ASR). For reference, we present automatic speech recognition performance on a set of the databases using the Bell Labs ASR system. For the databases with good recording conditions, the word-error rates can be less than 1%. In order to promote speech science and technology for real world applications, we make this database available for the speech community.

1. INTRODUCTION

Automatic connected digit speech recognition is essential for speech recognition research and for many telephone-based applications. Such applications include credit card and account number validation, catalog ordering, and digit dialing using voice. The high accuracy on connected digit string recognition is critical to the success of many real-world applications that require accurately recognizing phone numbers, account numbers, currency and other numbers during the transaction process.

In 1994, we started our effort to standardize Bell Labs training and testing procedures on connected digits speech databases. The objective was to provide a realistic benchmark for reporting connected digit performances and to facilitate the studies into specific problems, such as false start, extraneous speech input, background noise, endpoint detection, etc. To organize and to verify this largely expanded connected digit database inventory, there was such a need to develop an approach that takes advantage of the developed ASR system to reduce the manual work in improving the quality of the databases.

In order to adequately test connected digit speech recognition systems, we have assembled databases representing the ranges and complexity of conditions that will be occurred in real applications. These databases range in scope from talkers that read prepared lists of digit strings to customers that use an ASR system to access information regarding their credit card accounts. Data was collected over both analog and digital lines using a variety of telephone handsets.

The BLCD databases also included a collection of digit strings that have been recorded under different environmental conditions. The training set of the BLCD combine digit strings from various networks (analog, digital, and mixed) and collection conditions (pre-defined digit string read-out and real service field trials). A challenging testing set is defined which includes spoken digit strings from a collection of non-overlapped speech data collection.

2. BLCD DATABASES

In the BLCD databases, all data were recorded over the landline telephone networks. The digit strings were recorded under two kinds of conditions: Speakers read digit strings from a predefined list; or speakers talked spontaneously in real-world telephone applications. This section briefly describes each of the seven databases.

2.1. 1988 Mall Database (Mall88)

The Mall88 database was recorded in shopping malls across 22 dialectally distinct regions, which coincide with the regions used by Texas Instruments in collecting the TI Connected Digits Database within the United States (e.g. [4]). The specifications of this database are listed as follows:

- 100 talkers (50 males and 50 females) per region.
- The ages of the speakers are from 18 to 70.
- 110 digit strings per speaker with the length from one to seven digits. The contents of the digit strings were provided in the form of written text.
- Analog phone line terminated at an analog or digital central office (CO), and the toll connecting trunk was analog, digital, or mixed
- At each data collection end, an analog local loop was connected to the C.O. with the data collection system which convert analog voice signal to 16-bit PCM format.
- Four telephone handsets were used: two electrets and two carbons.

In the final release of the BLCD, only 15 sites of the Mall88 data collections were used in order to balance with the data from other collections. They are Long Island, NY (01), Pittsburg, PA (02), Dallas, TX (04), Chicago, IL (05), Boston, MA (06), New Orleans, LA (08), Miami, FL (09), Kansas City, KS (10), Denver, CO (12), Columbus, OH (15), Memphis, TN (18), Richmond, VA (19), Philadelphia, PA (20), Atlanta, GA (21), and Detroit, MI (22).

2.2. 1991 Mall Database (Mall91)

This database was recorded in shopping malls across 4 dialectally distinct regions within the United States (e.g., [2]). The specifications of the Mall91 database are listed as follows:

- 250 speakers (125 males and 125 females) per region.
- The ages of the speakers are from 18 to 70.
- Long digit strings, in the form of telephone numbers and credit card account numbers ranging from 10 to 16 digits in length, were provided in the form of a written text for reading.
- Digital recording over a T-1 interface.
- Four telephone handsets were used: two electrets and two carbons button.

The dialect regions of this data collection are Long Island, NY (02), Boston, MA (03), Birmingham, AL (04), and Chicago, IL (05).

2.3. Teletravel Database

The Teletravel database includes spontaneously spoken digit strings. Each talker spoke over different telephone lines using different telephone handsets. The entire connected digits database consists of about 3000 speakers. Each one of them pronounced his/her ten-digit telephone number. All data were recorded digitally.

2.4. UCS Database

The UCS database was collected as part of a field trial of connected digit recognition. Customers desiring to retrieve their current account balances were required to speak their 16-digit card number along with their five-digit zip code. The database included 3496 16-digit credit card numbers only.

Each number was spoken by a different person. An analog tip-ring connection was used in the recording set-up.

2.5. OSPS Phase I Database

The OSPS database was collected in Bloomington, Indiana, using an IVR platform. This database represents the first phase of a field trial of connected digit recognition for increasing operator automation of calling cards and third-number calls. Each speaker spoke either a ten-digit telephone number or a fourteen-digit credit card number. Two test sets were defined, namely, 960 calling card numbers (card), and 804 telephone numbers (tel). The digit strings were collected from speakers that provided "expected" responses to system prompts with ten digits for phone numbers and 14 digits for calling card numbers.

2.6. VoiceCard Database

The VoiceCard database was collected in a field trial. Digit strings were digitally recorded over a T-1 interface. About 200 speakers were required to speak their account numbers and pin numbers. Digit strings that ranged between 1 and 10 digits in length were collected using different microphones and cellular phones. In this study, only a subset of this database including about 4500 strings was used.

2.7. VoicePro Database

The VoicePro database consists of over 2000 isolated digit strings (plus some non-digit words). The originally digital data were passed through an analog leg and recorded on the Network Services Complex (NSCX), an ISDN digital telephone network system.

3. THE BLCD ORGANIZATION

We divided the BLCD databases into two sets: training set and test set. Only a selected number of digit strings was used from each database. This is due to the vast amount of data available. The reason for this selection is to provide a realistic benchmark for reporting connected digit performance within a reasonable time frame and without being biased by a specific database. A uniformly balanced random utterance selection method was used.

3.1. Training Set

The training set includes both read and spontaneous digit input from a variety of network channels, microphones and dialect regions. It consists of spoken digit strings from the Mall88, Mall91, Teletravel and UCS databases. Two subsets for training were defined: clean and exceptions. This dichotomy is useful in constructing good digit models as well as models of clicks, background noise, breathe, etc. The clean set consists of a subset of the four databases, namely, 15 sites of the Mall88, four sites of the Mall91, Teletravel and UCS. The data distributions of the training set are listed as follows:

- Mall88: 50 speakers per site approximately; 13-19 strings per speaker.
- Mall91: 125 speakers per site; 1-5 strings per speaker.
- Teletravel: 2075 speakers per site; one string per speaker.
- UCS: 2639 speakers per site; one string per speaker.

The exceptions included spoken digit strings from the four training databases. Table 1 shows the numbers of digit strings available in the clean and exception sets for training.

3.2. Test Set

The test set is designed to have data strings from both matched and mismatched environmental conditions. Currently, it includes the following databases: Mall88, Mall91, Teletravel, UCS, VoicePro, VoiceCard, and OSPS Phase I. Spoken digit strings affected by an AT&T TrueVoice^{sm} channel are also included in two of the databases, namely, VoicePro and Teletravel. Overall, the testing database forms a challenge to many ASR systems and is a vehicle for investigating the different characteristics of each database.

The testing data include two subsets as well:

clean and exceptions. The data split is to facilitate the study of some specific problems, such as false starts, disfluency, etc. It also helps evaluate different techniques for robustness and rejection.

Database	Training		Testing	
	Clean	Exceptions	Clean	Exceptions
Mall88 (site01)	503	73	514	93
Mall88 (site02)	578	160	576	146
Mall88 (site04)	574	173	583	151
Mall88 (site05)	609	94	664	88
Mall88 (site06)	983	25	615	174
Mall88 (site08)	525	69	529	329
Mall88 (site09)	578	133	546	195
Mall88 (site10)	404	174	502	145
Mall88 (site12)	577	213	579	212
Mall88 (site15)	531	85	552	358
Mall88 (site18)	511	0	516	0
Mall88 (site19)	651	120	663	191
Mall88 (site20)	525	167	539	167
Mall88 (site21)	625	188	601	158
Mall88 (site22)	631	297	632	93
Mall91 (site02)	647	241	651	271
Mall91 (site03)	642	85	672	65
Mall91 (site04)	608	191	619	100
Mall91 (site05)	671	88	707	54
Teletravel	2075	264	518	264
Teletravel(tv)	0	0	518	0
UCS	2639	268	713	106
OSPS (card)	0	0	792	168
OSPS (tel)	0	0	543	261
VoiceCard	0	0	3063	1268
VoicePro	0	0	2159	513
VoicePro(tv)	0	0	2159	513
Total	16087	3108	21725	6083

Table 1: Distribution of spoken digit strings among the training and test sets of the BLCD database

The clean set consists of a subset of the seven databases, namely, Mall88 (15 sites), Mall91 94 sites), Teletravel, UCS, VoiceCard, VoicePro and OSPS. The testing sets for the Teletravel and VoicePro databases were passed through an AT&T TrueVoice^{sm} channel. These are referred to as Teletravel(tv) and VoicePro(tv). The OSPS Phase I database has two subsets: calling card number (card) and telephone numbers (tel). The final data distributions of the testing set is as follows:

- Mall88: 50speakers per site, 13-19 strings per speaker.
- Mall91: 125 speakers per site, 1-5 strings per speaker.
- Teletravel: 518 speakers, one string per speaker
- UCS: 713 speakers, one string per speaker.
- VoiceCard: 200 speakers, 5-15 strings per speaker.
- VoicePro: 50 speakers, 5-15 strings per speaker.
- OSPS: 1281 speakers, one string per speaker.

The exceptions included digit strings from all the seven databases. The number of digit strings used per database is shown in Table 1.

4. DATA VERIFICATION

Although most of the databases were manually verified elsewhere, further verification was necessary and it was performed with the following procedure (see Figure 1).

The objectives of the verification procedure are two folds:

1. Verify labeling, and correct labeling errors in the string lexicon.
2. Divide each of the training and the testing sets into two categories, namely, “clean" and “exceptions." Exceptions are files with certain defects, such as clicks, touch tones, etc. This was necessary for the investigation of specific problems in connected digits recognition.

Files that contained extraneous acoustical events other than the expected digit string were classified as exceptions. For the BLCD databases, about 19.3% of the training set and 28% of the testing set were categorized as exceptions. These amounts are believed to be sufficient for the research and evaluation of different voice verification technologies [4].

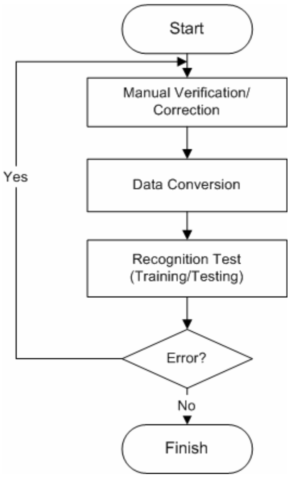
Start
Manual Verification/ Correction
Data Conversion
Recognition Test (Training/Testing)
Error?
Yes
No
Finish

Figure 1: The BLCD Database Verification Procedure

5. ANNOTATION AND TRANSCRIPTION VERIFICATION

In the BLCD database, the annotation, transcription, and recording conditions that associated with the speech are stored in a Bell Labs developed speech file header [4]. (It is named SSW header, a standard speech file header within the Lucent, AT&T, and Avaya speech community.).

Only the "Mallxx" databases have speaker gender and age range annotations. Other field-trial databases do not have these annotations.

```
ATT_SSW
246
database "universal card"
coding mulaw dsp
collect mode service
language "American English"
vocabulary "digits"
tel interface tip ring
file "univ.0113/pool0.0/268738050ac1"
label 4480 51440 "5 3 9 8 5 5 9 Z Z 1Z 8 \ 3
Z 9 Z" acnt bkns
```

Figure 2: An Exception SSW Header Example

For the exception files, the following label codes were used to identify the exceptions:

ng	bad file
sil	no audible sounds in file
bkns	background noise
stat	static
hum	hum
bksp	background speech (by someone other than speaker).
nnkw	non_keyword sound inforeground
nksp	non_keyword sound by subject
nspn	non_speech noise
lgh	laugh
brth	breath
cgh	cough
clk	click
tt	touch-tone
onhk	handset placed on hook
cut	keyword was cut off
h2h	hard to hear key word
misp	mispronounced
acnt	speaker with accent
wnkw	wrong number of key words
wkw	wrong key word

6. SPEECH RECOGNITION RESULT

In this section, we present the testing results on a set of the BLCD databases as references. In this task, the LPC coefficients (LPCC) with short-term energy were used as the speech feature. The acoustic model trained by discriminative training algorithms consists of 1-silence model and 275 head-body-tail, context dependent sub word digit models. The databases have been described as above, and they all contain pure digit strings. In all the evaluations, a recently developed endpoint detection algorithm [6] was applied. Both endpoint detection and energy normalization were performed in real-time mode, and only the detected speech portions of an utterance

Databases	Number of Strings	Number of Words	Word Error Rates
Mall88 (site02)	576	1738	1.1%
Mall88 (site04)	583	1743	1.5%
Mall88 (site05)	664	2087	0.7%
Mall91 (site02)	619	8194	0.7%
Mall91 (site04)	651	8452	5.6%
Mall91 (site05)	707	9426	1.4%

Table 2. Speech Recognition Results

were sent to a real-time recognition back-end. Models and parameters for endpoint detection were unchanged throughout the evaluation in all databases to show the robustness of the algorithm. The evaluation results are listed in Table 2.

7. DISCUSSIONS

The BLCD databases have a broad landline telephone network environment coverage that can be used for robust speech recognition for both research and product developments. Although it can be used for general voice network speech technology development, including wireless voice network, it does not cover a wide range of the fast growing digital wireless voice network environments. To address the specific problems in wireless network, we collected another set of speech databases from digital wireless voice network. The wireless databases will be introduced in our future publications.

8. CONCLUSION

The BLCD databases have been used extensively within Lucent, AT&T and Avaya for speech research and development. It was used as the main research databases for robust connected digit recognition. Many Bell Labs' inventions and speech research results were tested and evaluated on these databases to validate their performances in real-world applications before deployments. It has been used in internal speech product development as well. To promote speech science and technology for real-world applications, Lucent now makes this database available to academic community for non-commercial speech research. It is available for commercial use licensing as well. Please contact the authors of this paper for more details regarding the distribution of this database.

The BLCD databases are available for academic and research use under Lucent non-exclusive, non-commercial, limited-use license [6]. BLCD are available for commercial licensing as well.

8. ACKNOWLEDGEMENTS

The authors would like to acknowledge all our current and former Lucent, AT&T, Avaya, and Agere speech community colleagues those who took part in the data collection and verification of the BLCD databases. Special thanks to Mazin Rahim, Vivan Karam, Lathi Toni, and those who spend the entire summer 1994 to make the 1122CD database (the former name of BLCD) available. We also thank Dan Krasinski, Greg Hoke, Jeff Webb, Eric Buhrke, Caterina Maxwell, Margo Tikijian, Rich Sachs, Ben Stern, Bill Wetzel, Ze'ev Rivlin, Ann Syrdel, Wu Chou, Ilija Zeljkovic and Karen Chu, and colleagues of Lucent Speech Solution group. Without their efforts, these large, high quality speech databases were impossible to be collected and constructed.

References

[1] Rahim, M; Karam, V; Toni, L; Wilpon, J; Zhou, Q., "1122CD – A Collection of Telephone Based Connected Digits Speech Databases: Release Version 1.0," Lucent ITD-95-25724K, 1993.

[2] Wilpon, J. G., Buhrke, E., Chou, W., Juang, B.-H., Lee, C.-H., Rabiner, L., Rahim, M., Rivlin, Z., and Zeljkovic, I., "Connected Digit Recognition for Telecommunication-based Applications," Lucent ITD-94-20921P, 1993.

[3] Sachs, R., Stern, B. J., and Wetzel, W. R. "Speech File Header Standard for Speech Input Technology: Issue 1," Lucent ITD-93-13982Y, 1993.

[4] Leonard, R. G., "A Database for Speaker-Independent Digit Recognition," Proc. ICASSP'84, Mar. 1984, pp. 4.2.11.1-4.

[5] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," IEEE Trans. on Speech and Audio Processing, March 2002.

[6] "Lucent Public and Non-Exclusive Limited-Use Software from Bell Labs," http://www.bell-labs.com/topic/swdist.