

# A DATABASE FOR SPEAKER-INDEPENDENT DIGIT RECOGNITION

R. Gary Leonard

Texas Instruments Incorporated  
Central Research Laboratories  
P.O. Box 226015, MS 238  
Dallas, Texas 75266, USA  
(214)995-0388

## ABSTRACT

A large speech database has been collected for use in designing and evaluating algorithms for speaker independent recognition of connected digit sequences. This dialect balanced database consists of more than 25 thousand digit sequences spoken by over 300 men, women, and children. The data were collected in a quiet environment and digitized at 20 KHz.

Formal human listening tests on this database provided certification of the labelling of the digit sequences, and also provided information about human recognition performance and the inherent recognizability of the data.

## DESCRIPTION OF SPEAKERS

The number of speakers contributing data for the database is 326. The number of speakers and the age range of the speakers for each of the categories Man, Woman, Boy, and Girl are shown in Table 1.

TABLE 1. Number and Age Ranges of Speakers

	Symbol	Number	Age Range
Man	M	111	21 - 70
Woman	W	114	17 - 59
Boy	B	50	6 - 14
Girl	G	51	8 - 15

In order to obtain a dialect balanced database, the continental U.S. was divided into 21 dialectical regions (1), and speakers were selected so that there were at least 5 adult male and 5 adult female speakers from each region. In addition, 5 adult black males and 6 adult black females were selected. There was no attempt to dialect balance the child speakers.

## VOCABULARY DEFINITION

The utterances collected from the speakers are digit sequences. Eleven digits were used: "zero", "one", "two", ..., "nine", and "oh". Seventy-seven sequences of these digits were collected from each speaker, and consisted of the following types.

- 22 isolated digits (two tokens/digit)
- 11 two-digit sequences
- 11 three-digit sequences
- 11 four-digit sequences
- 11 five-digit sequences
- 11 seven-digit sequences

Hence each speaker provided 253 digits and 176 digit transitions. A unique set of prompts was prepared for each speaker. The following algorithm was used to generate the list of prompts for a given speaker.

1) Generate a 77-element array containing the digit sequence lengths. (22 elements of this array were 1 and 11 each were 2, 3, 4, 5, and 7.) These 77 elements were then randomized uniformly, and used to determine the position in the prompt list of the sequences of a given length.

2) For the 22 isolated digits, randomly select (without replacement) from a list of two tokens of each of the eleven digits.

3) To determine the first digit in each of the 55 sequences of length 2 or more, randomly select (without replacement) from a list of 5 tokens of each of the eleven digits.

4) To determine succeeding digits in a sequence of length 2 or more, randomly select (without replacement) from the "transition list" corresponding to the previous digit. There are 11 transition lists, one corresponding to each of the 11 digits, and they initially contain 2 tokens of each of the 11 digits. Should more than 22 transition tokens from any of the transition lists be required to complete a sequence, then the entire procedure is begun again (go to step (1)).

To prevent "zero" and "oh" from both occurring in the same sequence, as soon as a "zero" or an "oh" is selected, the transition list for the other pronunciation is relabelled; i.e., for example, if a "zero" is selected first, then if an "oh" occurs, a "zero" is substituted.

This procedure makes the frequency distribution uniform over all eleven digits. However, the "zero"- "zero" and "oh"- "oh" transitions tend to occur twice as frequently as any other transition.

#### DATA COLLECTION

The speech data were collected during the period June-September, 1982. Each speaker was seated in an acoustically treated sound room (Tracoustics, Inc., Model RE-244B acoustic enclosure), with the microphone placed 2-4 inches in front of the speaker's mouth. The microphone was Electro-Voice RE-16 Dynamic Cardioid.

Using a semi-automatic interactive data collection utility program executed by a DEC VAX 11/780 computer, prompts (determined as in the previous section) were presented one at a time to the speaker using large characters on a VT100 CRT. The program monitors the incoming speech and uses an energy measure to determine utterance beginning and end. The utterances were digitized using a Digital Sound Corporation Model 200 16-bit A/D/A. The sampling rate was 20 KHz, and a 10 KHz anti-aliasing filter was used. The sampled speech data were stored in ILS-compatible files on disk, ready for immediate review and re-collection, should that be necessary.

The monitoring capability (looping the digitized data back through the D/A) of the DSC Model 200 allowed the data collector to hear the data being stored to disk, thereby providing positive aural verification of data integrity. The collection utility program monitors and displays to the collector the maximum signal level of collected utterances, and provides an immediate indication of possible failures of the automatic segmenter. These features made the data collection procedure very efficient in terms of time required for data collection and amount of high quality data collected.

#### CERTIFICATION OF THE DATA

In order to determine the actual speech content of the collected utterances, formal human listening and classification of the data was conducted. As a by-product of this data certification

we also studied (1) the performance of humans in recognizing digit sequences, and (2) the inherent recognizability of the data. The data which the listeners heard was an LPC synthesized version of the original data, so that we were also able to judge whether the LPC parameterization preserves sufficient information to allow high performance recognition.

#### METHOD

Twenty-six listeners were hired to take part in the listening tasks. Sixteen panels of three listeners each were formed. Each of the listeners on fifteen panels listened to the 1540 utterances obtained from a unique set of 20 different speakers, while the 16-th panel listened to the utterances obtained from the remaining 26 speakers.

Each utterance in the database was downsampled to 12.5 KHz, analyzed and synthesized using 14-th order autocorrelation LPC analysis. A 25 ms window length and 10 ms frame period were used with pre-emphasis constant of 0.9375. Pitch tracking was accomplished using a crosscorrelation algorithm with post-processing (2). Listeners heard only this synthesized speech.

Using an interactive listening utility program, and listening individually in a quiet environment, each listener keyed-in a digit sequence to indicate the sequence of digits heard in each utterance. To encourage accurate transcription, a listener's base pay of \$50 was augmented by a \$10 bonus should all sequences be transcribed correctly, and the base pay was reduced by \$1 for each sequence transcribed in error (allowing a minimum pay of \$25, however). To further facilitate accurate transcription, the listening utility was designed to minimize the effect of inattention and typographical errors. Each utterance could be heard on command as many times as desired, and a backup feature allowed correction of typing errors. Only the keystrokes corresponding to the digits 1,2,...,8,9,Z, and 0 were allowed as responses, thereby reducing keystroke errors. The listener responses were displayed on a CRT as large characters to allow visual verification of the keyed-in responses. The listeners were asked to complete their task within one week. Within this constraint they could complete as many utterances as they wished in a given listening session. This helped reduce errors due to fatigue.

When all three listeners on a panel had finished their task, the utterances they had classified were analyzed as follows. The response string keyed-in by

each listener was compared to the string of digits requested from the speaker for that utterance. If all four of these digit strings were exactly the same, then it was assumed that the speaker had uttered the requested digit string, and no further analysis was done. Otherwise the utterance was flagged, and more detailed analysis was carried out. This analysis included comparing the three listener's responses and careful listening to the LPC data as well as the original data sampled at 20 KHz. Of the 25,102 collected utterances, 136 were flagged for further analysis.

For 30 of the 136 flagged utterances, we found that the speech data obtained during data collection was not the requested sequence of digits. These resulted from speaker errors not detected by the data collector. There were, of course, many speaker errors which were detected, and for which reprompts provided correct utterances. The remaining 106 utterances were flagged due to errors committed by one or more of the listeners. Detailed analysis of the 136 flagged sequences follows.

#### SPEAKER ERRORS

The 30 speaker errors can be categorized into the following seven types: (1) Omission of a digit; (2) Insertion of a digit; (3) Transposition of two digits; (4) Substitution of another digit for the correct digit; (5) False start, followed by a correction; (6) Sequence spoken as a numeral (e.g., the sequence "4 2" was spoken as "forty-two"); and (7) Pause between digits too long, causing the segmenter to cease digitizing prematurely. The total number of speakers involved in these errors is 26; four speakers made two errors each. These 30 residual speaker errors are 0.12% of the total collection of utterances. Note that the errors are not indicative of actual speaker performance since the number of errors detected by the data collector is not available for inclusion in this analysis.

#### LISTENER ERRORS

Listeners committed 116 errors involving 107 flagged utterances. In 35 instances, a speaker keyed-in "0" instead of "Z" when hearing the digit "Zero". In 2 instances, the listener failed to press the rubout key before entering the correct response, thereby leaving original keystrokes as part of the recorded response. Putting aside these obviously operational errors, the remaining 79 errors may involve errors in perception and can be categorized into these four types: (1) Transposition of two digits;

(2) Insertion of non-existing digits; (3) Omission of a digit; and (4) Substitution of an incorrect digit for the correct digit. The utterances involving these 79 errors were spoken by 53 different speakers.

Since each of the 25,102 utterances were heard by three listeners, the occurrence of 79 listener errors implies that the per-utterance listener perceptual error was at most 0.105%. The number of digits involved in the speaker errors is 84. Since each of 326 speakers was requested to say 253 digits, the per-digit listener perceptual error is at most 0.034%.

#### INHERENT RECOGNIZABILITY OF THE DATA

In an effort to estimate the ultimate recognizability of the collected digit sequences, the performance of a "super-human" committee classifier was determined. The decision of this classifier was defined to be the majority of the decisions made by the three listeners on a panel.

There were 62 utterances for which only one of the three listeners on a panel committed a listening error. The committee decision was correct for these utterances. However there were 7 utterances for which two of the three panel members made identically wrong classifications, and there was one utterance for which all three panel members made identically wrong classifications. That is, the committee made errors on 8 utterances. Five speakers spoke the 8 utterances; three utterances from one speaker, two utterances from a second speaker, and one utterance each from three other speakers. There was one digit in error in each of these 8 utterances. Hence the ultimate recognizability of the (LPC synthesized) data was measured as 99.99%.

To determine whether human classification could be improved by allowing collaboration, 3 listeners (who had not been involved with this database) were asked to transcribe, at first individually, the 70 tokens which caused listener errors. Of these 70 tokens, there were 17 for which at least one new listener committed a classification error. Then these listeners were brought together for further consideration of the 17 tokens. They were allowed to hear these utterances upon command, and were asked to discuss their perceptions and arrive at one answer, designated the "consensus" decision. The result was that 7 utterances remained incorrectly classified, indicating that collaboration provided no significant improvement over

the "committee" decision.

To determine whether the orginal data digitized at 20 KHz could allow better human classification than the synthesized data, 3 additional listeners provided a consensus classification of the 70 tokens, using the original data as their input data. Individual classification yielded 7 of the 70 tokens misclassified, and collaboration reduced the number of misclassified tokens to 2. The committee decision rule yielded 1 token misclassified. Hence the original data allows correct perception in certain cases where the LPC representation does not.

#### DATABASE PREPARATION

To prepare the database for use, meaningful filenames were assigned to the data files, the data were divided into training and testing subsets, and the data were copied to digital magnetic tape.

First, 6 of the 30 utterances which contained speaker errors were deleted from the database. Two of these utterances contained non-digit speech, two utterances contained a truncated digit, one utterance contained both digits "oh" and "zero" as the result of a substitution, and one utterance contained eight digits as the result of an insertion. The remaining 24 utterances containing speaker errors were relabelled to indicate the actual digit sequence spoken. One of these relabelled sequences contains six digits, and is the only 6-digit sequence in the database.

The database was divided into two subsets, one to be used for algorithm design and one to be used only for evaluation. The division was based on speaker category (M,W,B,G) and dialect classification, and yielded two sets of speakers, each containing approximately half the speakers of each category and classification.

The filename assigned to each data file consists of 3 to 9 characters and is of the form "NSI". The symbol N represents a string of 1 to 7 of the characters Z,1,2,3,4,5,6,7,8,9,0 and indicates the spoken digit sequence. The symbol S represents a 2-letter speaker designator (initials). (The letters Z and O are not used in speaker designators.) The symbol I is either null or a single digit, and is used to distinguish multiple utterances of the same digit sequence by the same speaker. The absence of a digit indicates there is only one utterance of the digit sequence by the speaker, while the presence of a digit M, say, indicates the M-th utterance of the digit sequence by the speaker. For example, the filename

"23Z45MA" was assigned to the file containing the first or only utterance of the sequence "2 3 zero 4 5" by speaker designated "MA". The filename "ODF2" was assigned to the file containing the second utterance of the sequence "oh" by the speaker designated "DF".

The following information was stored in the header of each data file:

- 1) Speaker's name;
- 2) Two-character speaker designator
- 3) Speaker's age
- 4) Speaker's dialect classification
- 5) Speaker's category (M,W,B,G)
- 6) Speaker's subset (Train,Test)
- 7) Sequence of digits uttered

The data files were then written to 32 digital magnetic tapes at 6250 bpi in ANSI standard file format.

#### ACKNOWLEDGMENT

We gratefully acknowledge the assistance of Thomas B. Schalk in locating the speakers and collecting the speech data for this database.

#### REFERENCES

1. E. Shochet and D. Connolly, "An Investigation into the Effects of Dialectal Variation on Flight Plan Filing by Machine Recognition," Interim Report, FAA-RD-80-115, January, 1981.
2. B.G. Secrest and G.R. Doddington, "An Integrated Pitch Tracking Algorithm for Speech Systems," Proc. ICASSP, April, 1983.