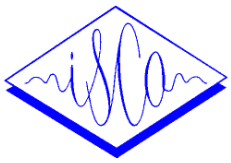

Speech Input/Output Assessment
and Speech Databases
Noordwijkerhout, the Netherlands
September 20-23, 1989

SPEECH DATABASE DEVELOPMENT: TIMIT AND BEYOND

Victor Zue, Stephanie Seneff, and James Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
U.S.A.

ABSTRACT

Automatic speech recognition by computers can provide the most natural and efficient method of communication between humans and computers. While in recent years high performance speech recognition systems are beginning to emerge from research institutions, scientists unequivocally agree that the deployment of speech recognition systems into realistic operating environments will require many hours of speech data to help us model the inherent variability in the speech signal. This paper describes the experiences of researchers at MIT in the collection of two large speech databases.

1. INTRODUCTION

Over the past five years, researchers at MIT have participated in several efforts devoted to the collection of speech databases. The development of databases is considered crucial because the acoustic realizations of phonemes depend on complex interactions among a multitude of factors. These factors can be *phonetic* (meaning that the realization of one phoneme may be severely affected by neighboring phonemes), *acoustic* (arising from changes in the acoustic environment and transducers), *intra-speaker* (due to variations in the talker's psychological and physiological state, speaking rate, and voice quality), and *inter-speaker* (due to differences in dialect, sociolinguistic background, and vocal tract size). A large body of speech data collected from many speakers will enable us to discover and quantify these context-dependent phenomena. Successful development of speaker-independent, phonetically-based speech recognition systems depends critically upon such data.

Speech databases can serve two additional functions: system training and system evaluation. Many present-day speech recognition systems achieve high performance by using powerful statistical models whose parameters must be estimated from an adequate amount of training data. Testing specific recognition algorithms or entire speech recognition systems on a common database provides a mechanism to compare their performances.

This paper describes the database development efforts in which we have participated, the most well-known being the TIMIT acoustic phonetic database. We will also report on a recent project directed towards collecting spontaneous speech elicited by simulating a realistic goal-oriented environment.

2. THE TIMIT ACOUSTIC-PHONETIC DATABASE

The TIMIT database design is the result of a joint effort among MIT, SRI, and TI. The corpus is comprised of 2342 distinct sentences from three different collections. First, two *calibration sentences,* provided by SRI, were designed to incorporate phonemes in contexts where significant dialectical differences are anticipated, and were spoken by all talkers. Next, 450 *phonetically compact* sentences were hand-designed by MIT with emphasis on as complete a coverage of phonetic pairs as is practical (Lamel et al., 1986). Each sentence was spoken by 7 talkers, in order to provide an indication for speaker variability. Finally, 1890 *randomly selected* sentences, chosen by TI, provide alternate contexts and multiple occurrences of the same phonetic

sequence in different word sequences. These were chosen primarily from the "Brown corpus" of American English sentences (Kucera & Francis, 1967), along with a few sentences from the Hultzen et al. corpus of playwrights' dialogue.

This combination of sentences was selected for its ability to balance the conflicting desires for compact phonetic coverage, contextual diversity, and speaker variability. It was decided by the research community that these three criteria were paramount to the initial acoustic-phonetic database. Each speaker read the two calibration sentences, five of the phonetically-compact sentences, and three of the randomly selected sentences, providing a total of ten sentences.

2.1 Data Collection

The recording of the data was carried out by researchers at TI (Fisher et al., 1986). A total of 6300 sentences were collected, ten from each of 630 speakers from eight dialectical regions of the United States. Approximately 70% of the speakers (439) are male and 30% are female. The speech data were digitally recorded at 20 kHz in a relatively quiet environment, simultaneously on a pressure-sensitive microphone and on a Sennheiser close-talking microphone. Digital tapes were shipped to NIST (formerly NBS), where they were filtered and downsampled to 16 kHz to prepare them for transcription processing at MIT.

2.2 Transcription

Each sentence in the TIMIT database has a time aligned transcription associated with the speech waveform. The transcription process consists of three steps. First, an acoustic phonetic sequence, obtained through listening and visual examination of various displays, is entered by a phonetician. Next, the acoustic phonetic sequence is automatically aligned with the waveform using the CASPAR system developed at MIT (Leung & Zue, 1984; Leung, 1985). Finally, the automatically generated boundaries are verified by experienced acoustic phoneticians, and manually corrected when necessary. Description of the inventory of acoustic phonetic units and the criteria for boundary placements can be found elsewhere (Zue & Seneff, 1988).

Once the phonetic transcription is aligned, it is rather straightforward to propagate the alignment up to the orthographic transcription as well as the intermediate phonemic transcription. A time-aligned orthographic transcription is useful when searching for a specific word, while time-aligned phonemic transcription can be used to relate the lexical representation of words to their acoustic realizations. The mapping of the time-aligned acoustic-phonetic transcription to the phonemic and orthographic transcriptions is accomplished using an automated procedure developed at MIT (Kassel, 1986).

2.3 Status

The acoustic-phonetic database was completely phonetically transcribed, aligned, and verified as of June 1988. As the sentences were completed, they were sent to NIST, where they were examined and prepared for distribution. The database was initially distributed to the general public via magnetic tapes. More recently, the first two-thirds of the TIMIT database has become available as a compact disc.

Many minor errors in the database have been found and corrected, both at MIT and at NIST, but despite our best intentions, more errors undoubtedly exist. It is our intention to continually provide corrections and updates for the foreseeable future.

3. THE VOYAGER SPOKEN LANGUAGE DATABASE

An effort in spoken language understanding has recently been initiated at MIT. The project is motivated by our belief that many of the applications suitable for human/machine interaction using speech typically involve interactive problem solving. That is, in addition to converting the speech signal to text, the computer must also *understand* the linguistic structure of a sentence in order to generate the correct response. In our initial efforts to develop a spoken language system, we have focused our attention on three main issues. First, the system must integrate speech recognition with natural language in order to achieve speech understanding. Second, the system must have a realistic application domain, and be able to translate spoken input into appropriate actions. Finally, the system must begin to deal with spontaneous speech, since people do not always utter grammatically well-formed sentences during a spoken dialog.

In order to explore issues related to a fully-interactive spoken language system, we have selected a task in which the system knows about the physical environment of a specific geographical area as well as certain objects inside this area, and can provide assistance on how to get from one location to another within this area. The system, which we call VOYAGER, currently focuses on the geographic area of the city of Cambridge, Massachusetts, between MIT and Harvard University.

VOYAGER is made up of three components. The first component, the SUMMIT speech recognition system (Zue et al., 1989; Zue et al., 1990), converts the speech signal into a set of word hypotheses. The natural language component, TINA (Seneff, 1989), provides a linguistic interpretation of the set of words. The parse tree generated by TINA is translated into a query language form, which is used to produce a response. Currently VOYAGER can generate responses in the form of text, graphics, and synthetic speech. It can deal with about a dozen distinct concepts including directions, distance and time of travel between objects, relationships such as "nearest," and simple properties such as phone numbers or types of food served. VOYAGER also has a limited amount of discourse knowledge which enables it to respond to queries such as: "How do I get *there*?" It can also deal with certain clarification fragments such as: "The bank in Harvard Square."

3.1 Data Collection

As a first attempt to create a spontaneous speech database, we have collected data from 50 male and 50 female subjects. Since the speech recognition part of VOYAGER is not running in near real time, data were collected under a simulation mode, in which spoken sentences were typed into the computer for automatic natural language processing and response generation. Each session, lasting approximately 30 minutes, started with the subject listening to a five minute introductory tape describing the task. The subject was then asked to conduct a spontaneous dialogue with VOYAGER, during which the voice input was typed into a computer log by an experimenter. (After the text was entered, the response time was one to two seconds on the average.) Following about 20 minutes of dialogue, the subject was asked to read their sentences from the computer log. Thus the database includes both a read version and a spontaneous version of (approximately) the same sentence, modulo false starts and filled pauses in the spontaneous version. This will enable a comparison to be made between spontaneous and read speech.

Both the user queries and the system responses were recorded on audio tape. The text log of the dialog included both the input sentences and the output generated by VOYAGER. Two of the sessions were recorded on video tape to document the data collection procedure.

3.2 Database Statistics

From the computer log, we were able to automatically generate some preliminary statistics of the resulting database. We have designated five male and five female speakers as the test set, with the remaining 90 speakers as the development set. Table 1 summarizes some of the relevant statistics for the development set. Note that the number of sentences refers to the spontaneous ones; the total number collected is double this amount.

As the table reveals, two-thirds of the sentences could be handled by the current version of VOYAGER. This means that one third of the data can be used to extend VOYAGER's capabilities. The remaining third consisted of equal amounts of out-of-vocabulary words and failed parses. Only a very small amount, about 1%, were parsed but not acted upon, which was the result of a conscious decision to constrain the coverage of the natural language component according to the capabilities of the back-end. While the number of unknown words appears to be large, they actually account for less than 3% of the total number of words when frequency of usage is considered.

The statistics of this database indicated that an average of slightly less than 50 sentences per subject were collected in each 20 minute dialogue. Thus we believe the database can easily be expanded as the capabilities of the system grow.

Table 1: Statistics on the VOYAGER Database

	Development Set
No. of Speakers	90
No. of Sentences	4361
Ave. No. of Words per Sentence	8.0
No. of Sentences with Action	2854(65%)
No. of Sentences with Unknown Words	740 (17%)
No. of Sentences with No Parse	727(17%)
No. of Sentences with No Action	40(1%)
No. of Words Used	601
No. of Unknown Words	398(66%)

4. DISCUSSION

While results in controlled environments are encouraging, the current capabilities of automatic speech recognizers are inadequate for general acceptance and use. In uncontrolled sites and situations, the error rates are one or two orders of magnitude too high. Almost surely the reason for this falloff in performance is inadequate statistical modelling of the variabilities of the kind mentioned at the beginning of this paper, and perhaps of kinds we have not yet considered.

This paper has described our experiences with two large speech databases, which were designed to meet somewhat complementary objectives. The TIMIT database was intended to be task and speaker-independent, and is suitable for general acoustic-phonetic research. The VOYAGER database, on the other hand, was intended for development and evaluation of a system which incorporates both speech and natural language processing. This database is particularly valuable as a source of spontaneous utterances elicited in a realistic goal-oriented environment.

While these databases constitute a great deal of speech data, there are strong indications that the data requirements of the speech community will soon be one or two orders of magnitude larger. In an attempt to meet these projected needs, a group of researchers from industry, universities, and government have recently proposed a national effort in the U.S. directed towards the collection and distribution of a large speech database. The database is to be used to aid in the development, training and evaluation of future speech recognition systems. A joint effort in database development will enable each individual site to obtain more data than they could alone, and can particularly aid sites who do not have much experience in collecting large databases.

ACKNOWLEDGEMENTS

The development of the databases described in this paper has benefitted from the help of many people, both within and outside of MIT. They include: Jared Bernstein, Corine Bickley, Bill Fisher, Jim Hieronymus, Katy Isaacs, Rob Kassel, Lori Lamel, Hong Leung, Dave Pallett, and Lydia Volaitis. This work was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

REFERENCES

Fisher, W. et al. (1986) "The DARPA Speech Recognition Research Database: Specifications and Status," *DARPA Speech Recognition Workshop Proceedings*, Texas Instruments Inc., Computer Sciences Center.

Kucera, H. & W.N. Francis, (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, R.I.

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

Leung, H. C., & V. W. Zue, (1984) "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," *Proc. ICASSP-84,* pp. 2.7.1–2.7.4.

Leung, H. C., (1985) "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Kassel, R. H., (1986) "Aids for the Design, Acquisition, and Use of Large Speech Databases," S.B. thesis, Massachusetts Institute of Technology.

Seneff, S., (1989) "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proceedings from ICASSP*:711-714, Glasgow, Scotland.

V. W. Zue, & S. Seneff (1988) "Transcription and Alignment of the TIMIT Database," *Proc. Second Meeting on Advanced Man-Machine Interface through Spoken Language,.*

Zue, V., J. Glass, M. Phillips, & S. Seneff, (1989) "Acoustic Segmentation and Phonetic Classification in the SUMMIT Speech Recognition System," *Proceedings from ICASSP*:389-392, Glasgow, Scotland.

Zue, V., J. Glass, M. Phillips, & S. Seneff, (1990) "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," Submitted for presentation at ICASSP-90.