Creation of Spoken Hebrew Databases

Tami Rannon, Ofra Golani, Anat Goren, Sherrie Shammass, Ami Moyal

NSC
Natural Speech Communication Ltd.
33 Lazarov St., P.O.Box 5212
Rishon-LeZion 75150
Israel
{tami_r, ofrag, anattr, shaunie, amym}@nsc.co.il

Abstract

Two Spoken Hebrew databases were collected over fixed telephone lines at NSC – Natural Speech Communication. Their creation was based on the SpeechDat model, and represents the first comprehensive spoken database in Modern Hebrew that can be successfully applied to the teleservices industry. The speakers are a representative sample of Israelis, based on sociolinguistic factors such as age, gender, years of education and country of origin. The database includes, digit sequences, natural numbers, money amounts, time expressions, dates, spelled words, application words and phrases for teleservices (e.g., call, save, play), phonetically rich words, phonetically rich sentences, and names. Both read speech and spontaneous speech were elicited.

1. Introduction

There is a growing demand for Automatic Speech Recognition (ASR) for a wide range of telephony applications. Service providers would like to implement ASR as an alternative to cumbersome and time-consuming phone keypad input for such applications as Voice Activated Dialing, Voice Controlled Voice Mail and Business Assistant. The incorporation of such services requires collecting a database (DB) of large amounts of recorded speech from a representative sample of speakers.

The DB should be statistically designed to cover all relevant speech variation, including speaker differences, dialect differences and coarticulation effects, and should be collected using a variety of telephone channels.

Hebrew, the official language of Israel, is spoken by over 5 million people. To date, there has not been speech recognition databases created for this language that could be useful in servicing the telecommunications industry. This paper outlines the development of two spoken Hebrew databases, representing a unique attempt to meet the speech recognition needs for the teleservice industries in the Hebrew language (Golani et. al, 1997; Cole et. al, 1995).

2. Description of Modern Hebrew Language

Modern Hebrew is unique in having been revived from Biblical Hebrew, a dormant language used mainly for religious rituals, to a daily spoken language. It exhibits influences from written forms of various historical periods. In use for a mere one hundred years, it is in a constant state of change. The main force of change is due to the huge impact of immense waves of immigration representing a multitude of languages (Kaddari, 1984). Hebrew belongs to the Semitic group of languages, which includes Arabic, Amharic, and Aramaic. It has a triconsonantal root structure, which is most clearly reflected in the syllabic writing system. Most related words can be traced to a common three-consonantal root structure, with the vowels prescribed by defined grammars. In the case of verbs, the vowels follow specific patterns that are called 'Binyanim', literally, 'houses'. Other word forms also have specific vowel patterns (Ben-Hayyim, 1972; Glinert, 1989).

Grammatically, Hebrew marks gender, person, time, and plurality. Morphological agreement is required between nouns, adjectives, and verbs. Thus, there are many morphological suffixes, which are semantically redundant within any given sentence. This is important to note since it affects the statistics of phoneme occurrence within the language. Hebrew grammar basically follows the order of Subject-Verb-Object, but the initial subject is often omitted and marked only morphologically in the verb. This illustrates the type of grammatical variation expected when eliciting spontaneous speech utterances. There are a total of 30 phonemes in Modern Hebrew, but some highly common Hebrew words use sounds from a variety of other languages. When the Hebrew language was revived, a great deal of attention was paid to the actual pronunciation of specific words, as well as updating a vocabulary that was lacking many modern concepts. Although it was originally intended to have only 'pure' Hebrew renditions for any given term, this proved an impossible task, and many foreign-based words slowly but surely crept into the language. Modern Hebrew is today composed of elements from various languages, such as Yiddish, Russian, Arabic, English, French, etc. Dialect variation in Israel is not geographically based, but rather, based on the country of origin. Though immigration to Israel is from all over the world, we usually consider two distinct variants of spoken Hebrew: Ashkenazic, stemming from Western Jews, and Sepharadic, spoken by Eastern Jews. Sepharadic pronunciation typically has guttural allophones, more similar in nature to Arabic. About 35% of Israelis are 'first generation' in Israel, meaning that their parents immigrated to Israel from another country, and about 35% more are not native Hebrew speakers.

Hebrew is written from right to left. Vowel sounds are generally not indicated, and mature readers fill in these missing vowel elements during the reading process. Beginners, however, are first taught to read and write with special vowel symbols (called "Nikud", or dots) and as their language skills mature, the vowels are removed (Kaddari, 1984; Morag, 1972).

Without vowel markings, ambiguities in reading can occur. For example, when vowel notation is not provided, the words "sefer" (book), "sapar" (hairdresser) and "safar" (counted) are written exactly the same – "spr". Disambiguation is done during the reading process by higher level semantic and syntactic decoding of the context. Such ambiguity is not present when the vowel symbols are provided.

3. Database Description

3.1 General Description

NSC has collected two spoken databases recorded over fixed networks using analogue lines. (Analogue lines are more common than digital lines in Israel). The databases are intended for training and testing several types of speech recognizers such as isolated word recognition, word-spotting, continuous speech and phoneme-based systems.

The database design was patterned after that of the Fixed Database (FDB) in the SpeechDat project in terms of both linguistic content and speaker sample (Höge et. al, 1997; Winski, 1997; Gibbon et. al, 1997). In contrast to other databases, geographical location of the speaker was not a factor, but rather, the country of origin. Thus, information regarding the speaker's origin, or the speaker's parents' origins was noted. Speakers were chosen from all over Israel. Only speakers who were able to read Hebrew fluently were allowed to participate.

Two methods were used for speaker recruitment:

- 1) Company employees enlisted relatives and friends.
- 2) An advertisement was put in the daily newspapers.

Several tens of utterances were elicited from each speaker, comprising a mixture of spontaneous and read speech. For read speech, vowel notations were supplied in order to eliminate any ambiguities that are common when using the standard non-vowel notation. Thus, 'standard Hebrew orthography', in this case, refers to the inclusion of the vowel symbols, even though they are usually omitted in printed material.

3.2 Database I (DB I)

DB I consists of 250 speakers, with each speaker providing 32 utterances, both read and spontaneous. The read corpus includes keywords, digits, dates, times, spellings, names for directory assistance and phonetically rich words. The spontaneous recorded material includes yes/no questions, names, digits, times, and dates.

3.3 Database II (DB II)

DB II consists of 850 speakers, with each speaker providing 75 utterances, both read and spontaneous. The read corpus includes teleservice command words, word spotting phrases with embedded commands, digits, natural numbers, money, dates, times, spellings, names for directory assistance and phonetically rich words and sentences. The spontaneous recorded material includes yes/no questions, names, digits, times, and dates.

4. Database Content

Since the DB I content is a subset of DB II, we describe the full content of DB II. Specific differences are clearly shown in the summary table below (Table 1).

4.1 Read Speech

Read Speech was elicited by having speakers read from a prompt sheet written in standard Hebrew orthography, with vowel notation. The prompt sheet included the following items:

Isolated Digits

Since Hebrew has both feminine and masculine forms of digits, both forms were elicited by including digit words written out in standard Hebrew orthography. The Hebrew words for 'hash' and 'star' were also included.

Isolated Digit Sequences

A randomized digit sequence with commas inserted after each digit was presented. The speaker was asked to read the digits separately, pausing after each digit.

Digit/Number Strings

The following types of digit strings were elicited: a unique prompt sheet number (nnnnnn), an Israeli telephone number (0n-nnnnnnn or 0nn-nnnnnn), a credit card number (nnnn-nnnn-nnnn-nnnn) and a Personal Identification Number (PIN) (nnnnnn). The speaker was explicitly asked to read the strings digit by digit, and not as natural numbers.

Natural Numbers

Natural numbers between ten to one million were elicited. Special care was taken to sample all units of numbers, such as tens, hundreds, thousands, etc.

Money Amounts

Various typical phrases used for money amounts were elicited, including both small and large money amounts, along with the appropriate terminology of Israeli currency. The term for 'American dollars' was also provided, since many large commodities such as apartments or cars are quoted in American dollars.

Times

Times can be expressed in either analogue or digital form. Phrases elicited included: day parts (morning, noon, afternoon, evening and night), time units (hour, minutes, seconds, o'clock), time parts (quarter to, quarter past, half past), relative time units (after, before), time approximations (exactly, approximately, nearly, about, etc.) and numbers representing the hour and the minutes. These phrases were a sample of phrases taken from expanded BNF grammars of possible forms.

Dates

Hebrew dates may be specified in both analogue and number form. The prompt sheets included only the analogue form, including Hebrew equivalents of weekdays, Gregorian months, and numbers representing 'date' and 'year'. In addition, relative date phrases were elicited (next month, yesterday, Jewish holidays, etc.). These phrases were a sample of phrases taken from expanded BNF grammars of possible forms.

Spelled Words

Hebrew speakers rarely resort to using spelling when handling misunderstandings of the intended message. However, it is still necessary to include random letter sequences and spelling of some names in order to train material suitable for recognizing acronyms or alphanumerical codes. The spelled items were random letters and places names. All off them were written with commas after each letter, and the speaker was asked to say them as spelling. There are no 'upper' or 'lower' cases in written Hebrew. Several words such as the Hebrew equivalents of 'hyphen', 'apostrophe', etc. were also included.

Application Keywords/Key-phrases

Since gender considerations would effectively double the list of keywords, half were in the masculine form, and half were in the feminine form. In Spoken Hebrew, the future form frequently replaces the imperative form. Thus, about 70 application-oriented keywords were needed in order to represent 25-teleservice functions (such as call, save, play, etc.). These words were presented in isolation and were also embedded in sentences to provide a basis for word-spotting tests.

Phonetically Rich Sentences

Each speaker read several sentences out of a list of 600 phonetically rich sentences. These sentences were devised in order to cover adequate training of monophones, and the most frequent biphones and triphones. They were of variable length, and included simple sentences as well as complex sentences. To provide prosodic variety, declaratives, imperatives and interrogative sentences were included. Standard Hebrew orthography, with vowel notations, was used in all cases.

Phonetically Rich Words

Each of the speakers read a few isolated words from a list of about 1000 distinct words covering at least 100 occurrences of each of the Hebrew phonemes. Some examples of 'foreign' words, which have entered the language and are highly common, were also included.

Standard Hebrew orthography, with vowel notations, was used in all cases.

Directory Assistance Names

In order to provide testing material for directory assistance, the prompt sheet contained one city name, one company name, and one proper full name.

4.2 Spontaneous speech

Spontaneous speech was elicited by asking the subjects specific questions. The following information was elicited.

Time

The speaker was asked for the current time, and could answer in either analogue or digital form.

Date

The speaker was asked for his birthday, and could answer in either analogue or digital form.

Spelled Words

Each speaker was asked to spell his forename.

Yes/No questions

Each speaker was asked 3 questions - one question that would probably elicit a YES answer, one that would probably yield a NO answer, and one that would probably elicit a 'fuzzy response', such as 'I don't know', 'Maybe', 'I'm not sure', etc.

Directory Assistance Names

In order to provide testing material for directory assistance, each speaker was asked for his name and his place of birth.

Digit Sequence

Each speaker was asked to say a familiar phone number. The speaker was explicitly asked to read the strings digit by digit, and not as natural numbers.

5. Recording, Annotation and Verification

A recording station was specially designed and built for the collection of the database. Each utterance was recorded on a separate file with an appropriate annotation file in SAM format. The VOX! Annotation software was used (CSELT, 1997).

All material was transcribed and annotated. An orthographic transcription of the speech actually pronounced by the speakers, along with the 'expected' utterances (i.e., what the speaker was supposed to say) was provided for each utterance. Orthographic labeling of the actual spoken utterances was done using SAMPA labeling conventions and SAMPA notation. The SAMPA annotation provides transliteration that resolves ambiguous pronunciation of Hebrew words. Clearly the correct Hebrew spelling is not apparent from this transliteration. The 'expected' utterances were written in

standard Hebrew orthography in order to make the transcription easier for the transcribers. Phonetically rich utterances (words and sentences) and names were both orthographically and phonetically annotated. All utterances were verified by having one transcriber check utterances that were originally transcribed by a different transcriber.

The following summary table shows the specific content of both databases.

Read Speech	DB I	DB II
Isolated Digits	3	2
Isolated Digit Sequences	-	1
Digit Strings - Read strings	4	3
Natural Numbers	-	3
Money Amounts	-	3
Times	1	1
<u>Dates</u>		
Jewish Holidays & Relative	-	5
General – Gregorian Date	1	4
Spelled Words	3	2
Application Keywords	6	16
Application Key-phrases	ı	9
Phonetically Rich Sentences	-	8
Phonetically Rich Words	3	5
Directory Assistance Names	3	3
Spontaneous speech		
Time	1	1
Date	1	1
Spelled Words	1	1
Yes/No questions	2	3
Directory Assistance Names	2	3
Digit Sequence	1	1
Total	32	75

Table 1: Specific Content of DB I and DB II

6. Future work

In order to support speaker independent applications, we intend to extend this database to a total of 5000 speakers. We plan on collecting a similar database over Israeli cellular networks since there is a very high penetration of cellular phones in the Israeli telecommunications market. A collection of an Arabic database is also planned since there are more than one million Arabic speakers in Israel proper, with many more Arabic speakers in the surrounding regions.

7. Summary

This paper represents a first attempt to collect Hebrew Spoken Databases useful for the teleservice industry. The collection and annotation of such databases was accomplished successfully. Validation of the two

databases that were collected ensured that they were of high quality. Part of the data was used as training material, and the recognition results were very good.

Following the SpeechDat guidelines proved to be very helpful. However, changes in the design were necessary to accommodate specific aspects of the Hebrew language, as well as sociolinguistic factors in Israeli society. The Hebrew test case shows that there is a need to combine specific knowledge of database creation with expert knowledge of the language and a thorough understanding of the local culture.

DB I will be ready for use at the second half of year 2000.

8. References

Zeev Ben-Hayyim, 1972, *Hebrew Grammar*, Encyclopaedia Judaica, Jerusalem, Vol. 8, pp. 77-102.

Ron Cole et. al., 1995, The challenge of Spoken Language Systems: Research Directions for the Nineties, IEEE Tran. on Speech and Audio Proc., Vol. 3, pp. 1-20.

CSELT, 1997, VOX! User Guide, Torino.

Dafydd Gibbon, Roger Moore, Richard Winski (Eds.), 1997, *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin New York.

Lewis Glinert, 1989, *The Grammar of Modern Hebrew*, Cambridge, Cambridge University Press.

Ofra Golani, Ami Moyal, Yaron Himmelhoch, Mordechay Grutman, Sherrie Shammass and Zeev M. Rannon, 1997, Subword based Name Recognition for Information Retrieval Applications, Proc. of the Int. Conf. on Speech Processing (ICSP), Seoul, Korea, pp. 469-473.

Harald Höge, Herbert S. Tropf, Richard Winski, Henk van den Heuvel, Reinhold Haeb-Umbach, Khalid Choukri, 1997, European Speech Databases for Telephone Applications, Proc. ICASSP'97, pp. 1771-1774

Menahem Zevi Kaddari, 1984, Introduction: The State of the Art of Israeli Hebrew. In: Ben-Zion Fischler (ed.) In: "Hebrew Books, Articles and Doctoral Theses on Contemporary Hebrew Published in Israel" (1948-1980). (From the Workshop: Studies and Research for Teachers of Hebrew as a Second Language, 6). Council on the Teaching of Hebrew (Hebrew), Jerusalem.

Shlomo Morag, 1972, *Proununciations of Hebrew*, Encyclopaedia Judaica, Jerusalem, Vol. 13, pp. 1120-1145.

Richard Winski, 1997, *Definition of Corpus, scripts and standards for Fixed Networks*, SpeechDat Technical Report LE2-4001-SD1.1.1, released on January 22.

Acknowledgements

Part of the project, DB I, was funded by the Israeli Ministry of Science, as Part of Call for proposals of 1997.