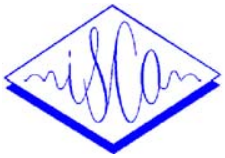

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

GREEK SPEECH DATABASE FOR CREATION OF VOICE DRIVEN TELESERVICES

I. Chatzi[1], *N. Fakotakis*[2], *G. Kokkinakis*[2]
[1]KNOWLEDGE. S.A., Human-Machine Communication Dept.
N.E.O. Patron-Athinon 37, 264 41 Patras, Greece
Tel: +30.61.452.820, Fax: +30.61.453.819, E-mail: echatzi@patra.hol.gr

[2]Wire Communications Laboratory (WCL), Electrical & Computer Engineering Dept.
University of Patras, 261 10 Patras, Greece.
Tel. +30 61 991 722, FAX: +30 61 991 855, E-mail: fakotaki@wcl.ee.upatras.gr

ABSTRACT

In this paper we present the collection of Greek speech data over the telephone network from 5,000 speakers in order to form a speech database (SpeechDatII.GR). This work is embedded in the Language Engineering Project LE2-4001 SpeechDat, in which all official European languages and some major dialectal variants are represented. The design of the speech database allows the development of word, phoneme and syllable based speech recognizers that can be used for a large variety of real speaker independent applications. In particular it will provide a realistic basis for training and assessment of both isolated and continuous speech recognizers for telephone speech, which is a prerequisite for developing voice driven teleservices.

1. INTRODUCTION

Telephone based services such as telebanking, telephone directory assistance, voice mail and time-table handling are evolving in the last few years as partly or fully automated, using modern speech technologies such as speech recognition, speech understanding and speaker recognition. The creation of spoken language resources for training and assessment of recognizers etc. is a prerequisite for the creation of voice driven teleservices.

Currently many automated teleservices rely on isolated word recognition, where words can only be spoken in isolation. Often only a single word is expected after a system prompt. However, in the near future the requirement for more user-friendly systems will arise, where:

- Continuous speech recognition (no pause between words required) will become prevalent.
- Command keywords can be embedded in a phrase, where the recognizer is able to extract the relevant command or keyword and discard the rest. This makes the service more robust and thus more user friendly.
- The user will be able to barge in, interrupt the announcement and not wait for the system prompt until he can give a command.
- Teleservices will evolve from a rigid machine-driven dialogue to a more natural user-driven dialogue. The user will be able to speak in a natural, spontaneous manner and provide/ask for information in any order, not being restricted to a prespecified menu structure.

In order to avoid a new data collection for every new application with a different vocabulary, we will work on algorithms for 'vocabulary independent' recognition: The new application vocabulary will be transcribed into sequences of phonemes (or other units/syllables), the models of which will be constructed from recordings of phonetically rich sentences. The intended data collection will thus encompas both application specific vocabulary, for those applications where the vocabulary is known (figures, usual commands etc.), and phonetically rich sentences, suitable for the compilation of new application vocabulary.

The following sections describe the work that is currently conducted for the development of SpeechDatII.GR, followed by some analytical results of the first 2,000 recordings and conclusions. Section 2 describes the distribution of the 5,000 speakers according to their region age and sex. In section 3, the database contents are shown, while in section 4 a small description of the recording platform is given. Sections 5, 6 and 7 present the orthographic transcription and lexicon conventions, as well as the CD-ROM SpeechDatII.GR directory and file structure.

2. ENVIRONMENTAL AND SPEAKER SPECIFIC COVERAGE

The telephone speech material is collected directly over the telephone network (EURO-ISDN). The training data for the Greek database comprise five thousand (5,000) speakers. The speech samples are gathered from speakers of various age, both male and female (≈50% males and ≈50% females) with a good representation of the major regional dialects of the Greek language. The age distribution, in order to cover the corresponding voice variation, is as shown in Table 1.

Age	Middle Point	Minimum %
0 - 7	4	0%
8 - 15	12	1%
16 - 30	23	20%
31 - 45	38	20%
46 - 60	53	20%
61 - ...	78	0%

Table 1. *The age distribution for the 5,000 speakers of SpeechDatII.GR.*

The regional background of the speakers has large effects on their speech. People speak differently depending on the specific region in which they grow up. The areas of the Greek territory that have been defined according to pronunciation variations are [2]:

1. *Standard (Urban) Modern Greek,* that covers Athens (Attica), South Euboia, Thessaloniki, the Peloponnese, Kythera and the Ionian islands. This is the language spoken and written by most Greeks in both formal and informal discourse. 81% of the recorded speakers will be from this area.
2. *Northern and Semi-Northern Modern Greek,* that covers Lefkas, Sterea Hellas, Ipiros, Thessalia, Macedonia, Thrace, North Sporades, Thasos, Lemnos, Imvros, Lesbos, Samos and Tinos. 10% of the recorded speakers will be from these areas.
3. *Cretan*, which is the dialect spoken in Crete. 6% of the recorded speakers will origin from Crete.
4. *Aegean Modern Greek* that covers the Dodecanese, the Cyclades, south Sporades and Chios. Only 3% of the recorded speakers will be from these areas.

The number of speakers to be selected per region is proportional to the region's population, presented in Table 2.

No.	Region	Population	Target	
1	Standard	7.750.000	4050	81%
2	Northern & Semi-Northern	1.000.000	500	10%
3	Cretan	600.000	300	6%
4	Aegean	300.000	150	3%
Total		**9.650.000**	**5000**	

Table 2. *Number of speakers selected per region.*

Another issue that is considered is the calling environment of the speakers. Calls can be made from telephone booths, office, home, factories, public places, street, vehicles, etc. We have decided that a minimum of 2% of calls will be made from public places with high background noise. In addition, each caller will inform us of the handset type that she/he is using (rotary, touch-tone, cordless,...) and the type of network (PBX, Fixed).

All this information on environmental and speaker specific coverage will be reported in the database description files.

3. DATABASE CONTENTS

The specifications of the contents give a total of 56 utterances per call comprising a mixture of spontaneous and read speech. This results in call durations of approximately 12-14 minutes each. The definition of the corpus contents is shown in Table 3.

Nr.	Utterance Description
2	**isolated digit items :**
1	single isolated digit
1	sequence of 10 isolated digits in one utterance
4	**digit/number strings :**
1	prompt sheet number (6 digits)
1	telephone number (9-11 digits)
1	credit card number (14-16 digits)
1	PIN code (6 digits)
3	**natural numbers :**
2	numbers (1-million)
1	decimal number
1	**money amount**
2	**yes/no questions :**
1	predominantly yes including "fuzzy" yes/no (spontaneous)
1	predominantly no including "fuzzy" yes/no (spontaneous)
3	**dates :**
1	birthdate (spontaneous)
1	prompted date phrase
1	relative and general date expression
2	**times :**
1	time of day (spontaneous)
1	prompted time phrase, in analogue form
3	**application keywords/keyphrases**
1	**word spotting phrase using embedded application words**
5	**directory assistance names :**
1	city of birth/growing up (spontaneous)
1	most frequent cities (set of 500)
1	most frequent companies/agencies (set of 500)
1	proper name (forename and surname)
1	proper name (set of 500)
3	**spellings :**
1	real/artificial word
1	spelling e.g. of directory assistance city name
1	spelling of proper name (spontaneous)
4	**phonetically rich words**
9	**phonetically rich sentences**
3	**syllabifications**
=45	**UTTERANCES / CORPUS**
+11	general purpose questions
56	**TOTAL utterances**

Table 3. *Specification of the contents of SpeechDatII.GR.*

4. THE RECORDING PLATFORM

The recording platform is a Pentium PC (100MHz, 16Mbytes RAM, 4G Hard Disk), running under Windows NT. The Line Interface Card that is used is the AVM ISDN Controller B1 that covers simultaneously 2 channels. The Telephone Link is the Basic Rate Euro-ISDN (ISDN-BRI). The programming Interface is CAPI 2.0 and the Application Software is written in C++. The Backup Software of Windows NT is being used.

Several evaluation and validation tests have been made on the recording platform. The validation methodology has been based on the following steps:

- *Expert Test*: It consists of four different tests: a "Listening Test" on recorded items, an "Overload Test" by loading the platform with multiple calls in parallel, a "Crash Test" by simulating interruption of power supply, and a "Stability Test" to check the failures frequency of the platform.
- *Functional Test*: This test has been performed by ten naive callers. Each of them had to call the platform and complete the dialogue by reading or answering the questions of the prompt sheet. Then two questionnaires had to be filled: one concerning the comprehension of the instruction and prompt sheet, and one concerning the dialogue quality.

5. ORTHOGRAPHIC TRANSCRIPTION

The transcription of the speech corpora is an orthographic, lexical transcription with a few details included that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. Extra marks contained in the transcription aid in interpreting the text form of the utterance.

The software EMPHASIS is used as a tool for processing digital signals [9]. This has been extended to include speech annotation functions. It enables the creation of an annotation file for each speech file. This annotation file includes information about:

a. graphic transcription of the speech signal,
b. phonemic transcription of the speech signal using a predefined phonetic table,
c. discrimination of the speech signal into speech and silence intervals,
d. location of the endpoints for each segment included in the speech signal,
e. correction of the already existing annotation files after listening to recorded speech files.

6. SPECIFICATION OF THE SPEECH DATABASE FORMAT

After the transcription and the validation of the speech corpora, the database will be stored on CD-ROMs. These disks will be printed according to the ISO 9660 Interchange level 1 specifications. In addition to speech files (presented as couples of label and signal files) data files will be supplied, including: a speaker demographic file, a pronunciation dictionary and a full database contents file.

The directory structure that is being followed is that of EUROM0 and EUROM1. This structure matches the user needs in terms of speech material availability. Its main characteristic is that it is content independent and easily built in the recording time. No further modifications are needed during the database processing. A shallow directory nesting with contiguous numbers is used to identify the individual sub-directories and call directories. The four level directory structure that has been chosen is presented in detail in Table 4.

\<corpus>\<database>\<block>\<session>

<corpus>	**SpeechDatII.GR**
<database>	Defined as <name><#><language code> where: <name> = **FIXED** (Fixed telephone network) <#> = **1** for SpeechDatII project <language code> = **EL** (ISO 6392-letter code)
<block>	Defined as: **BLOCK<nn>** where <nn> is a progressive number from 00 to 99. Block numbers are unique and there will be as many blocks as needed to fill a CDROM. Theses numbers must be the first 2 digits used in <nnnn> described below.
<session>	Defined as: **SES<nnnn>** where <nnnn> is a progressive number from 0000 to 9999, that is the numeric call identification number.

Table 4. *SpeechDatII.GR Directory Structure.*

Both signal files and label files have to be put in the terminal node subdirectories. Since there are no more than 60 utterances per call, the total number of speech files and associated transcription files does not exceed the CD-ROM recommended limit of approximately 120 items in a directory. In addition, the directories that will be used to store the other types of files (documentation files, etc.) are shown in Table 5.

7. CURRENT STATE OF THE RECORDINGS

Up-to-now approximately 2,000 callers have been recorded. The recording process goes on smoothly and we are in the procedure of recruiting the next 3,000 speakers. The estimated time of the completion of the recordings is the end of August 1997. Further 3 months will be needed for the transcription and the development of the CD-ROMs. Analytical results of the progress are presented below:

- 92% of the recorded calls have been accepted. The remaining 8% have been discarded for reasons as:
  a. not all items being read,
  b. two much noise present in the recordings,
  c. two many errors made while reading the prompt sheet.
- 54% of the callers have been females and 46% males.
- The age distribution has been as follows: 76% between 16-30 years, 13% between 31-45, 8% between 46-60 and 3% over 61 years of age.
- 45% of the telephones used have been analogue and 55% digital. 7% of them have been cordless and 3% public card phones.
- The calling location for 75% of the callers has been their home, for 22% their office while 3% called from a public place.
- The social status of the callers has been defined according to their occupation. 15% of the callers belong to the Upper Middle Class (Higher Managerial class), 23% to the Middle Class (Intermediate Managerial), 18% to the Lower Middle Class (Junior Managerial), 17% to the Skilled Working Class (Skilled Manual Workers), 22% to the Working Class (Semi and Unskilled manual Workers) and 5% to the lowest level of subsistence (unemployed persons, etc.).
- 67% of the callers have been smokers and 33% non smokers.
- 96% have a Greek nationality and 4% have a mixed or a foreign nationality.
- 73% of the callers came from the Standard Modern Greek region mainly from the areas of Athens and Patras, 12% from the Northern Greek territory, 2% from Crete, 2% from the Aegean and the rest 2% from abroad (Canada, Australia, USA).

8. CONCLUSION

We have described various components of the SpeechDatII.GR speech database and discussed some of the issues at its design. We believe that the speaker distribution, as far as their age or region is concerned, gives an adequate coverage of the Greek population. The contents of the database provide a sufficient coverage of all phonemes, di-phones and tri-phones of the Greek language. The recording platform has been well established and evaluated. The transcription of the speech corpora is done by the use of the transcription tool EMPHASIS, and the processed speech files are stored in the SpeechDatII.GR CD-ROM.

The need for the multilingual SpeechDat II database is threefold: first, to acquire acoustic-phonetic knowledge for phonetic recognition; second, to provide speech for training recognizers for speaker independent applications; and third, to provide a common test base for the evaluation of recognizers.

\(root)	a *"readme"* file reporting the first description of the database, a *"volume identification"* file and a *"copyright"* file
\<database>\DOC	documentation
\<database>\TABLE	speaker, recording condition and lexicon tables
\<database>\INDEX	index files, e.g. contents file, corpus contents files, speaker list files,...
\<database>\PROMPT	prompt sheet if present (with appropriate sub-directory structure if needed);
\<database>\SOURCE	any source code supplied

Table 5. *SpeechDatII.GR Non -speech data files directory structure.*

9. REFERENCES

[1] Robert Browning, "The Greek Language, Mediaeval and Modern." Ed. Papademas, Athens, 1988.

[2] N.G. Kontosopoulos, "Dialects and Idioms of Modern Greek." Athens 1994.

[3] J. Zeiliger, "Publishing CD-ROMS from EUROM-1", doc. Ref. SAM-A/ICP/004/ V1, from Esprit project 6819 (SAM-A), SAM-A Periodic Progress Report, Year1, 1993.

[4] Francesco Senia, "Environmental and speaker specific coverage for Fixed Networks", LE2-4001-SD1.2.1 Technical Report, July 1996.

[5] Richard Winski, "Definition of corpus, scripts and standards for Fixed Networks", LE2-4001-SD1.1.1 Technical Report, July 1996

[6] Andrei Constantinescu, "Recommendations and Specifications of Annotation Tools", LE2-4001-SD1.3.3 Technical Report, January 1997

[7] Irene Chatzi, "Validation of the Recording Platform", LE2-4001-SD2.4 Technical Report, February 1997.

[8] Irene Chatzi, "Installation of the recording device and documentation", LE2-4001-SD2.1 Technical Report, October 1996.

[9] KNOWLEDGE S.A., "EMPHASIS User Manual", March 1996.

[10] URL: http://www.phonetik.uni-muenchen.de/ SpeechDat.html

[11] URL: http://www.wcl.ee.upatras.gr/