Development of the Estonian SpeechDat-like Database

Einar Meister, Jürgen Lasn, Lya Meister

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics at Tallinn Technical University
einar@ioc.ee

Abstract

A new database project has been launched in Estonia last year. It aims the collection of telephone speech from a large number of speakers for speech and speaker recognition purposes. Up to 2000 speakers are expected to participate in recordings. SpeechDat databases, especially Finnish SpeechDat, have been chosen as a prototype for the Estonian database. It means that principles of corpus design, file formats, recording and labelling methods implemented by the SpeechDat consortium will be followed as closely as possible. The paper is a progress report of the project.

1. Introduction

Estonian is a Finno-Ugric language spoken by almost one million people. About 90% of them live in Estonia; considerable amounts of speakers are located in Canada, Sweden, the USA and Russia. Estonian language has the status of the state language of Estonia, whereas 67.3% of inhabitants (ca 921,000 persons) speak it as the mother tongue.

Rapid development towards the Information Society has made personal computer a part of our daily life at work and at home. We need computers to process, maintain and exchange information, which in most cases exists in a linguistic form. A major amount of information available on Internet is in English, most of the computer software we are using daily is in English, too.

The development efforts of the human-computer interaction (HCI) during past few decades have been directed towards natural communication using spoken language input and output. Enormous effort has been put into research and development of automatic speech recognition technology by a large number of laboratories and companies all over the world. As a result, limited communication with computers using spoken language is today a reality for several languages, i.e. for languages with a large number of speakers, like English, French, German, Spanish, Italian, Japanese, Chinese, etc. What about languages with a small number of speakers, like Estonian? Will they survive in the information society?

According to experts' view, the only way for small languages to survive is development of the human language technology (HLT) including tools for both spoken and written language processing. Therefore, development of the HLT for Estonian plays a crucial role for the future of the language.

During last years the demand for the speech recognition in Estonian has been increased. The companies providing mobile and call-center services have shown especially active interest. The typical DTMF-based services are coming to be exhausted, for the development of new and more user-friendly services speech recognition technology plays a crucial role. The first attempts to implement small vocabulary word recognition for mobile parking system has been carried out a few years ago [1], lately another prototype for recognition of isolated Estonian digits has been proposed [2].

Today's ASR-technology is developed for non-agglutinative languages, Estonian as an agglutinative language, needs a different approach – the specific problems of speech recognition in Estonian have been discussed in [3]. Whatever technological approaches are implemented, a speech database is inevitably necessary for training and testing of all ASR-systems. Currently, no large Estonian speech databases developed for ASR needs do exist – in order to fill this gap the Estonian SpeechDat project has been launched.

2. Choosing a prototype

A number of different audio file formats, computer readable phonetic alphabets, and database standards do exist [4]. Many speech databases for different languages do exist (see http://www.icp.inpg.fr/ELRA/ or http://www.ldc.upenn.edu). One of the most frequently cited and exploited in ASR-development is the family of SpeechDat databases [5]. All together 20 speech databases based on common principles of corpus design, file formats, recording and labelling methods and validation criteria have been collected in 14 European countries.

Due to the well-established design principles the SpeechDat databases have been chosen to serve as a prototype for our project. A special attention is paid to the Finnish SpeechDat [6] as the closest prototype for the Estonian SpeechDat database.

3. Estonian SpeechDat

The goal of the Estonian SpeechDat project is to collect speech samples from a large number of speakers for speech and speaker recognition purposes. Duration of the project has been planned for 24 months divided into four main stages:

- Preparatory activities (9 months)
- Recordings (4-6 months)
- Segmentation and labeling (6-10 months)
- Completion (4-6 months)

The project is financed by the Ministry of Culture and Ministry of Education and is supported by EMT, the biggest mobile operator in Estonia.

3.1. Main characteristics and structure

The main characteristics of the Estonian SpeechDat database will be as follows:

- Sampling rate: 8 kHz
- Signal format: 8-bit A-law, mono
- Signal source: calls from fixed and cellular phones
- Calling environment: home or office
- Speakers: at least 2000 (1000 female, 1000 male)
- Speech items: isolated and connected digits, natural numbers, money amounts, spelled words, time phrases, date phrases, yes/no questions, names, application words and phrases, phonetically rich words and sentences.

The database will have three subsets:

- F-corpus – at least 1000 speakers from fixed network, one call per speaker;
- M-corpus – at least 1000 speakers from mobile network, one call per speaker;
- SV-corpus – at least 200 speakers from fixed and mobile networks, at least 10 calls per speaker.

The technical environment includes two recording workstations connected to two ISDN-lines via Basic Rate Interface. The possibility to call a toll-free number has been enabled. The automatic recording process is controlled by ADA-software (developed by Universitat Politècnica de Catalunya, Spain). The recording dialogue in Estonian assists the caller through the recording session. The signals from the recording workstations will be transferred into server for further processing. Three workstations with labelling software WWW Transcriber [7] have been set up, as well.

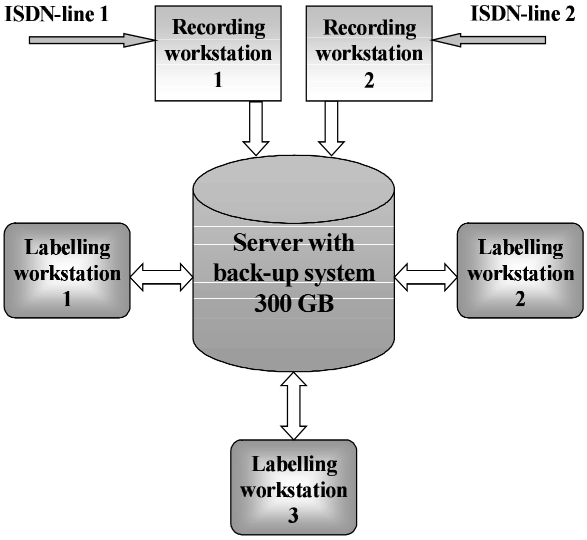

Figure 1. Structure of the local network for recording, storing and labelling of speech signals.

3.2. Corpus design

The SpeechDat databases have been designed to train several special-purpose speech recognisers, for example recognition of isolated command words, digit strings, numbers, dates and continuous speech, as well. The possible applications include different voice driven teleservices accessible via fixed and cellular network. The specific speaker verification corpus includes speech items recorded by the same speakers at different time intervals.

The Estonian corpus has been compiled according to SpeechDat design and includes:

- PIN-codes – randomly generated 6-digit strings
- Isolated digit string – 10 isolated digits in random order, all digits occur in a string only once
- Sheet number it is a five digit running number starting from 50000
- Telephone number – the format of telephone numbers is (0)xx-yyy yyyy, whereas (0)xx is the area code and yyy yyyy is the realistic telephone number. The area codes are real and the telephone numbers are generated randomly. GSM numbers are included, as well
- Credit card numbers – in the format xxxx xxxx xxxx xxxC. The Luhn checksum algorithm is implemented
- Time phrases – prompted either in digital format e.g. “21:35” or in analogue format “quarter past five”.
- Spontaneous time – spontaneous answer to the question "What time is it now?"
- Relative and general date expressions – the relative and general date expressions are drawn from a set of

the most usual expressions, like "tomorrow, yesterday, last Friday, the day after, etc."

- City names (local) – the list includes 150 names on cities, counties and villages in Estonia
- City names (foreign) – the list includes 100 names of larger cities and countries of the world
- Spelled items – are drawn from the lists of city names, person names and phonetically rich words
- Money amounts – prompted in the format x,y currency, in which x is a number between 0-100000 and y is a number drawn from the set of 0,5,...,95. The currency is chosen from the list of 10 most often used currency units in Estonia
- Person names – the list includes the most often 100 male and 100 female names in Estonia
- Company names – the list includes the most frequent local names as well as several foreign names
- Application words and phrases – includes most frequent commands of a typical IVR-service and different commands or menu items from the Estonian versions of MS Windows XP, Word and Excel
- Phonetically rich words and sentences – the list has been compiled taking into account the specific features of the phonological system of the language.

3.3. Speaker recruitment

Up to 2000 callers are expected to participate in the database collection. The speakers participating in recordings should fulfil several requirements related to gender, age, socio-economic factors, regional and dialectal factors, and environment-specific characteristics [8].

The demographic criteria to ensure a good coverage of the speaker population are as follows [9]:

- 50% (± 5%) male and female speakers,
- all accent regions have to be covered proportionally,
- distribution by age groups:
  - 16-30 years – min. 20% of speakers,
  - 31-45 years – min. 20% of speakers,
  - 46-60 years – min. 15% of speakers.

According to the experiences of SpeechDat projects the recruitment of speakers has been the most difficult and time-consuming task [10]. Different recruitment schemes – market research company, hierarchical, snowball, direct mail, newspaper and radio advertisements and calls, WWW-based recruitment – have been proposed [8].

We have used the following methods:

- calls for participation in radio, TV and newspapers,
- WWW-based recruitment,
- distribution of information via Public Access Internet Points (PAIPs),
- distribution of information at universities and companies.

All volunteers had to register by filling a special electronic questionnaire on the project's web-site http://www.phon.ioc.ee/base. Information collected from each speaker includes data about age, education, occupation, mother tongue, dialectal area, smoking habit, phone type, and contact address. After registration the recording instructions with free telephone number and the individual prompt sheet have been delivered, mostly by e-mail.

The actual number of registered speakers is displayed and regularly updated on the project's web-site. The preliminary target was to record 1000 speakers, but due to the successful recruitment schemes the new target – 2000 speakers – could be achieved to the end of 2003.

3.4. Comparison of recruitment schemes

Radio, TV and newspaper calls for participation were carried out in several times. Articles in several newspapers and live interviews in popular TV and radio programs have resulted in a substantial number of contributors.

WWW-based recruitment turned out to be the most effective way. A pop-up banner with the link to the project's web-site on the Internet-version of the largest newspaper in Estonia has attracted many readers to visit the project web-site. The banner advertisement has been repeated twice – at the beginning of the recruitment period (week 2) and three month later (week 15), and in both rounds the number of registered speakers increased by more than 500 volunteers (see Figure 2).

Distribution of calls for participation via Public Access Internet Points (PAIPs). There are about 500 PAIPs around Estonia, located mostly in public libraries and schools. PAIPs are actively visited by local inhabitants as in most places the access to Internet is free of charge. Each PAIP is managed by a local manager. The calls for participation were delivered to all PAIP-managers around Estonia with the request to distribute the information among their visitors. Most of the managers reacted with enthusiasm and as a result, speakers from all dialectal regions are participating in recordings.

Distribution of calls for participation in universities and companies. The calls for participation were distributed in main Estonian universities and in several companies. The results of these actions were not as successful as expected.

4. Current status

Currently (March 24, 2003), 1756 volunteers have been registered for the recordings. The number of successive calls is 837. Figure 2 shows the dynamics of recruited speakers and completed calls. Age distribution of contributors is given on the Figure 3.

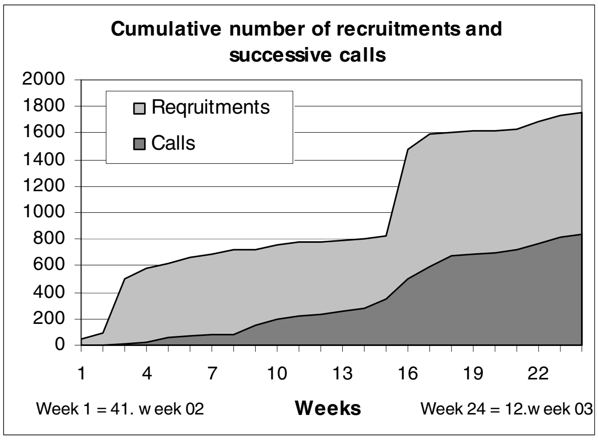

Figure 2. Increase in recruitments and in calls.

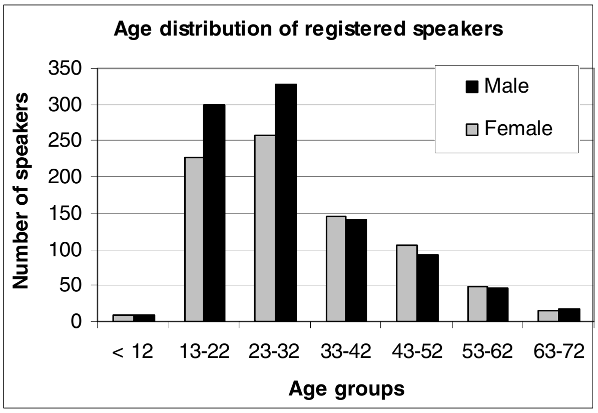

Figure 3. Age distribution of registered speakers.

5. Conclusion

The first year of the project has been successful – technical environment for recordings has been established, text corpus has been compiled, several speaker recruitment schemes have been implemented, and active recording period is going on. The WWW-based recruitment method turned out to be the most effective one. The preliminary target – 1000 calls – will be achieved within a few weeks and the final target – 2000 calls – should be achieved to the end of 2003.

Acknowledgements

The project is financed by the Ministry of Culture and Ministry of Education and Science, and supported by several institutions (EMT, Radiolinja Eesti, Tehnokratt, Vaata Maailma, Postimees, Õpetajate Leht).

The Estonian SpeechDat team is very thankful for help and valuable advice to the SpeechDat-expert Chistoph Draxler (University of Munich) and to Finnish SpeechDat team (Digital Media Institute, Tampere University of Technology), as well as to the expert on speech databases Toomas Altosaar (Acoustics Lab, Helsinki University of Technology).

All contributors are gratefully acknowledged, as well.

References

[1] Meister, E. et al. 2001. Spoken Dialogue System for Mobile Parking. Proceedings of the International Workshop SPEECH and COMPUTER, Moscow, Russia, 29-31 October, 2001. pp. 123-126.

[2] Alumäe, T. 2002. Eestikeelse kõne tuvastus: prototüübi loomine. Master thesis. Tallinn Technical University.

[3] Meister, E. 2001. Towards speech recognition in Estonian. Publications of the Department of Finnish and General Linguistics of the University of Turku (eds. S.Ojala, J.Tuomainen), 21. Fonetiikan Päivät, Turku 4.-5.1.2001. pp. 59-70.

[4] Altosaar, T. 2001. *Object-based modelling for representing and processing speech corpora.* Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing. Espoo. 98 p.

[5] http://www.speechdat.org

[6] Rosti, A. et al. 1998. *SpeechDat Finnish Database for the fixed telephone network.* SpeechDat Technical Report.

[7] Draxler, Ch. 1997. WWWTranscribe – A Modular Transcription System Base on the WWW. Proceedings of Eurospeech'97, Rhodos. Volume 4, pp.1691 – 1694.

[8] Lindberg, B., Comeyne, R., Draxler, Ch., Senia, F. 1998. Speaker recruitment methods and speaker coverage – experiences from a large multilingual speech database collection. Proceedings of ICSLP 98, pp. 2731-2734.

[9] Senia, F. et al. 1997. *Environmental and Speaker Specific Coverage for Fixed Networks.* SpeechDat Technical Report SD 2.2.1.

[10] Moreno, A. et al. 1998. Recruiting Speakers and Documentation on Speaker Typology. SpeechDat Technical Report SD 2.2.1.