SPEECON – Speech Databases for Consumer Devices: Database Specification and Validation

Dorota Iskra[1], Beate Grosskopf[2], Krzysztof Marasek[3], Henk van den Heuvel[1], Frank Diehl[4], Andreas Kiessling[5]

[1] Speech Processing Expertise Centre (SPEX), Nijmegen, the Netherlands (d.iskra, h.v.d.heuvel@let.kun.nl)
[2]Philips Speech Processing, Aachen, Germany (beate.grosskopf@philips.com)
[3]Sony, Stuttgart, Germany (marasek@sony.de)
[4]TEMIC Sprachverarbeitung GmbH, Ulm, Germany (frank.diehl@temic-sp.com)
[5]Ericsson Eurolab, Nuremberg, Germany (andreas.kiessling@eed.ericsson.se)

Abstract

SPEECON (Speech-Driven Interfaces for Consumer Devices) is a project which aims to develop voice-driven interfaces for consumer applications. Led by an industrial consortium, the project's goal is to collect speech data for at least 20 languages and 600 speakers per language (mostly adults but children as well). Recorded in different environments which are expected to be representative for the future applications, the database corpus comprises both spontaneous and read speech, the latter including phonetically rich material, a large number of application commands and isolated items such as digits, names, etc. In order to safeguard consistency and high quality of the databases, all of them are subject to validation. This paper describes in detail the specifications of the databases as well as the validation procedure.

1. Introduction

Speech-Driven Interfaces for Consumer Devices, otherwise known as SPEECON, is a shared-cost project funded by the European Commission under Human Language Technologies which is part of the Information Society Technologies programme (IST-1999-10003). The purpose of the project which started in February 2000 and will last till May 2003 is to develop voice-driven interfaces for consumer applications such as television sets, video recorders, mobile phones, palmtop computers, car navigation kits, information kiosks, and toys. Instead of operating these devices manually, users of these applications will simply need to talk to their equipment. The basic prerequisite for voice-driven interfaces is a collection of speech data for at least 20 languages (or regional dialects) including most of the languages spoken in Europe. At the same time special attention is paid to the environment of the recordings which represent typical surroundings of consumer electronics applications such as home, office, public places or moving vehicles (Siemund et al., 2000). The research activities of the project aim to develop algorithms for cross-environmental database adaptation (Gedge et al., 2002).

The goal of this paper is to present an extensive description of the specifications of the databases including corpus, speakers and recording environments. The second part of this paper focuses on the validation of the databases, or the quality checks, describing their purpose and procedure. Finally, a report on the current status of the project is given.

2. Project consortium

The SPEECON consortium comprises a number of industrial partners with Siemens as the project coordinator. The remaining partners are, in alphabetical order, Daimler Chrysler, Ericsson, IBM, Nokia, NSC, Philips, ScanSoft (formerly L&H), Sony, TEMIC Sprachverarbeitung. Additionally, Panasonic participates as an external partner. The appeal of the project is such that new internal and external partners are still joining the consortium and negotiations are currently underway with two new partners. Most of the partners have experience with earlier speech database projects of the SpeechDat family (Winsky et al., 1997).

3. Languages

Each partner, except for Daimler Chrysler, is responsible for data collection of two languages. Table 1 presents a list of all the languages collected in this project.

PARTNER	LANGUAGE	REGION
Siemens	Spanish	Spain
	Russian	Russia
Ericsson	Italian	Italy
	Swedish	Sweden & Finland
IBM	German	Germany & Austria
	English	United Kingdom
ScanSoft	Danish	Denmark
	Flemish	Belgium
NSC	Hebrew	Israel
	French	France
Nokia	Finnish	Finland
	Mandarin	P.R. China (incl. Hong Kong)
Philips	Dutch	The Netherlands
	Japanese	Japan
Sony	Polish	Poland

TEMIC	Portuguese	Portugal
	German	Switzerland
	English	USA
Panasonic	Spanish	USA
	Mandarin	Taiwan

Table 1: Language list

With the new partners the language set will be extended to include Hungarian, Czech, Cantonese and Thai.

For each language between four and six dialect areas are represented in the collection. For some of the languages where not enough dialectal variation is found, exceptions are made. The dialect of the speakers is based on where they grew up rather than where they were born or where they currently live.

4. Speaker specifications

For each language data are collected for 600 speakers: 550 adult and 50 child speakers. Among adult speakers 50% must be male and 50% female with a possible deviation of 5% (Kiessling et al., 2001). As gender-based speech differences between children are not evident, no gender restrictions are made for child speakers.

Child speakers are defined as toy users and roughly below the age of 15 (or the age of voice change for boys). This has implications for the corpus design of the child part of the database which includes a set of toy commands. Older children are seen as potential users of 'adult' applications such as mobile phones or information kiosks and form part of the adult database.

Three different age groups are distinguished for adult speakers (from 15, or the age of voice change, to 30-year-old, 31 to 45, 46-year-old and above) with at least 30% of all the speakers required to fall into each of the first two groups and 10% into the last. For children two age groups are defined: 8 to 10-year-old, 11 to 15-year-old (or the age of voice change). Again at least 30% of the speakers are required to fall into each subgroup. Table 2 summarises the speaker conditions.

	CHILDREN					
Age	8-10			11-voice change (15)		
%Required	≥ 30			≥ 30		
Gender	Male	Female		Male	Female	
%Required	no restrictions					
	ADULTS					
Age	15-30		31-45		46+	
%Required	≥ 30		≥ 30		≥ 10	
Gender	M	F	M	F	M	F
%Required	45-55	45-55	45-55	45-55	45-55	45-55

Table 2: Speaker specifications

Ideally, speakers should be naïve users rather than trained speakers, which should guarantee a wide range of speaking styles, voice qualities and regional influences, thus making the collection as representative as possible.

Except for Mandarin and US Spanish, the languages for which the concept of native speakers is hard to define, for all the speakers a particular language is required to be their mother tongue.

5. Recording scenarios

5.1. Recording environments

Consumer electronics comprise a wide range of varied applications and each of these can be used in one or more specific environments, e.g., an information kiosk is usually placed in a public place which may become noisy at times, a PDA (Personal Digital Assistant) can be used in quiet home or office conditions, but also on a noisy train or at an airport. On the basis of such application scenarios for the different devices, a number of acoustic environments were defined (Diehl et al., 2001).

- *Office*: mostly quiet; if background noise is present, it is usually more or less stationary.
- *Entertainment*: a home environment but noisier than office; the noise is more coloured and non-stationary; it may contain music and other voices.
- *Public place*: may be indoor or outdoor; noise levels are hard to predict.
- *Car*: a medium to high noise level is expected of both stationary (engine) and instantaneous nature (wipers).
- *Children*: handled as a separate environment, although similar characteristics to those of an entertainment environment are expected.

Speech data are recorded in all these acoustic environments in order to provide realistic training and test material. However, the emphasis is placed on the office and public place environments. Table 3 presents speaker distribution over all the environments. A deviation of 5% is allowed in each environment.

ENVIRONMENT	# SPEAKERS
Office	200
Public places	200
Entertainment	75
Car	75
Children	50

Table 3: Speaker distribution per environment

Additionally to the regular speech recordings for each speaker a sample of the actual average noise level in dB(A) is measured. If the recording equipment is set up to a new position, steps are also taken to estimate the reverberation characteristics in retrospect. These measures are obtained by a simultaneous recording of special reference signals (e.g., pink noise played back by a broadband loudspeaker) with two identical high quality omni-directional microphones under various predefined conditions.

5.2. Hardware and software set-up

The speech signal is recorded directly and simultaneously in four channels through four microphones which are mounted at different distances from the speaker's mouth:

- *close distance*: the microphones are a headset simulating the mobile phone position and a lavalier which is used when calling hands free and positioned just below the chin of the speaker. The speech signal of the headset recordings is expected to be of sufficiently high signal-to-noise ratio in order to serve as 'clean' signal suitable for training.
- *medium distance* (0.5-1m): characteristic of PDA, information kiosk and automotive applications; the speaker is face to face with the device; standard cardioid microphones are used here and an omni-directional measurement microphone is used specifically for the children and public place recordings;
- *far distance*: here an omni-directional microphone is used; the recordings with this microphone position are only made in office, entertainment and children environment.

The exact microphone arrangement depends on a particular scenario. Figure 1 presents a simplified equipment set-up (Diehl et al., 2001). The speech signal is recorded in raw format, sampled at 16 kHz and quantised using 16-bit linear coding.

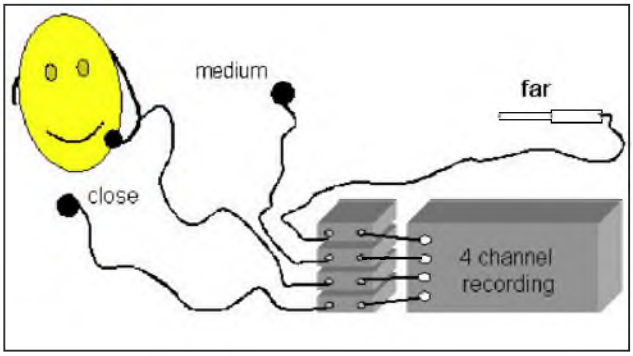

Figure 1: Equipment set-up

Apart from a laptop, all the partners use identical recording platforms consisting of a set of microphones, preamplifiers and other audio devices. Furthermore, the same recording software is used. Both these measures are meant to guarantee high quality and maximum consistency of all the databases.

6. Corpus and vocabulary specification

The specification of the content of the databases is similar to the SpeechDat projects in that possible application and interest areas were identified first and then respective corpus items and vocabulary were defined (Winsky et al., 1997; Höge et al., 1999; van den Heuvel et al., 1999; Siemund, 2000). The databases are a mixture of spontaneous and read speech, but also of continuous and isolated utterances (Marasek et al., 2001). SPEECON differs from its SpeechDat predecessors in that it contains spontaneous speech recordings and an extensive number of application specific commands.

6.1. Spontaneous speech

Spontaneous speech is recorded for adult speakers only and comprises about 10 spontaneous utterances per speaker derived from a list of 30 categories. The speakers are asked to respond to task specific situations related to mobile phone and PDA applications, automotive and information kiosks, audio/video and toys. For instance, in the navigation category speakers may be asked to describe the way to their favourite restaurant. Furthermore, a set of 17 questions are asked to elicit spontaneous responses containing times, dates, city and proper names, spellings, yes/no answers, languages and telephone numbers.

6.2. Read speech

A large part of read items are formed by phonetically rich material which is recorded in order to obtain sufficient training examples of all monophones and most frequent biphones and triphones in a language. This kind of material must enable continuous speech modelling using subwords. The phonetically rich part comprises a set of sentences and words of which every adult speaker utters 30 and five respectively. Every child reads 60 sentences.

The read part of the database contains also 31 general words and phrases containing isolated digits and strings, natural numbers, telephone numbers, money amounts, time and date expressions, spelt words, names of people, cities and streets, yes/no answers, e-mail and web addresses, and special keyboard characters.

The last category of the read part of the database comprises a set of application specific commands: about 500 for adult speakers and 120 for children. These were chosen to cover a number of different functionalities: basic IVR (Interactive Voice Response) commands, directory navigation, editing, output control, messaging and Internet browsing, organizer functions, routing, automotive, and audio/video. Application specific commands for children contain a set of general commands, toy commands and phone commands.

Table 4 presents a complete list of corpus items and the number of items uttered by each speaker. The number of items is chosen in such a way so that the session duration does not exceed one hour.

CORPUS ITEM CATEGORY	ITEMS PER SPEAKER
Spontaneous speech (adult speakers only)	
Free spontaneous speech: - mobile phone and PDA - automotive and information kiosk - audio and video - miscellaneous	10
Elicited speech:	
- dates: birth, current and relative	3
- time of day	2
- city names	2
- proper names	3
- spelt proper name	1
- yes / no answers	2
- language	1
- telephone numbers	3
Read speech	
Phonetically rich material:	
- sentences (60 for children)	30
- words (only adults)	5

General purpose words and phrases:	
- isolated digits	4
- isolated digit sequence	1
- continuous digit strings	4
- telephone number	1
- natural numbers	3
- money amount	1
- times: analogue and digital	2
- dates: analogue, digital and general	3
- spelt words	3
- personal name	1
- city / street names	2
- yes / no	2
- e-mail / web addresses	2
- special keyboard characters	2
Application specific commands (adults):	
- basic IVR	85
- directory navigation	40
- editing	22
- output control	57
- messaging and Internet browsing	70
- organizer functions	33
- routing	39
- automotive	12
- audio and video	95
Application specific commands (children):	
- toy commands	74
- general commands	34
- phone commands	14

Table 4: List of corpus items

The exact number of application specific commands can vary up to 10% per language if extra synonyms are added or discarded as a result of the translation. Each speaker reads only a part of the whole set of commands which is randomly chosen so as to provide at least 220 repetitions of each application command in the database.

7. Database validation

Validation is aimed at safeguarding the quality of the databases and their compliance with the specifications (van den Heuvel, 2000). Apart from the internal validations by the producers themselves, an external validation centre, the Speech Processing Expertise Centre (SPEX) performs the validation checks. These checks refer to the following aspects of the database:

- documentation
- completeness of the database
- file formats
- signal quality
- transcription quality
- lexicon
- speaker and environment distribution.

The databases are checked against a set of validation criteria which are derived from the specifications and assigned certain tolerance margins (van den Heuvel et al., 2001). In order to spot errors and inconsistencies at the earliest possible stage, validation is done in a number of phases.

7.1. Prevalidation

Prevalidation is the first phase and can be subdivided into two parts. The first part is constituted by all the checks which can be carried out before the actual recordings start. These checks concern the prompt sheets and the lexicon.

7.1.1. Prompt sheet and lexicon validation

Before the recordings start, all the prompt sheets are delivered to SPEX who check if all the corpus items are present; if no vocabulary items have been lost in the translation process as compared to the original specifications; if there is a sufficient number of repetitions of various vocabulary items; and if there is a sufficient number of phone repetitions for phonetically rich material. Naturally all these checks are done at the prompt level and have to be carried out again at a later stage at the transcription level when the recordings have been completed.

Together with the prompt sheets a lexicon is also submitted for validation. The lexicon contains all the orthographic entries together with their broad phonemic transcriptions according to the SAMPA protocol (Wells, 2002). Other information such as frequency of occurrence of words in the database, morpheme and stress information is optional. The formal correctness of the lexicon is checked by SPEX and concerns the format, the use of predefined SAMPA symbols and the consistency with the vocabulary content of the prompt sheets.

The quality of the phonemic transcriptions is evaluated by a network of phonetic experts who are native speakers of a particular language. They receive a random subset of the lexicon and their comments are fed back to the producers of the database.

7.1.2. 10-speaker database validation

In the second part of the prevalidation the recordings of the first 10 speakers are evaluated in order to find systematic errors at an early stage of the speech collection. For these 10 speakers identical checks are carried out as will be the case later for the complete database. These checks are executed on the speech files, label files, and documentation files and refer to the aspects enumerated in section 7. This validation phase is particularly appreciated by the producers of the databases as it allows them to spot, for instance, wrong hardware or software set-ups and make necessary adjustments before the recordings enter an advanced stage.

7.2. Complete database validation

This is naturally the most crucial part of the database validation. All the checks which were executed in the prevalidation phase (except for lexicon validation by phonetic experts) are carried out again; this time, however, on a complete database. Furthermore, orthographic transcriptions are evaluated by native speakers and the database is checked against a number of distribution criteria, such as gender or environment distributions, which is only possible when all the database recordings are available. The results of the validation are reported to the consortium who, in case of errors, evaluate their severeness and may reject a database or recommend necessary improvements. If these improvements are

substantial, the consortium may decide on a revalidation of the database.

7.3. Pre-release validation

In this last part of the validation a number of quick checks are performed in order to ascertain that all the most recent data are distributed on disks according to the specifications and that no data are missing. An exchange of the databases takes place only when the databases are reported to be of satisfactory quality.

8. Current status

The specification phase of the project was completed several months ago and all the partners are currently making recordings. With regard to validation, the prevalidation stage is finished and all the partners have submitted their prompt sheets, lexicons and 10-speaker databases. Most of the partners have already carried out a large part of the recordings as well and the first complete databases are expected to be submitted for validation at the time of the conference.

At the same time new partners are still joining the project with the commitment to complete their databases within the current time span. The project is, therefore, successfully on its way to make a major contribution to the currently available commercial speech resources.

9. References

Diehl, F., Fischer, V., Kiessling, A., Marasek, K., 2001. Specification of Databases – Specification of Recording Scenarios. Technical report nr D212 available from www.speecon.com.

Gedge O., Couvreur C., Linhard K., Shammas S., Moyal A. 2002. Speech Database Adaptation Methods for Speech Recognition at Cross-Environmental Conditions. *Proceedings of LREC 2002.*

Höge, H., Draxler, C., van den Heuvel, H., Johansen, F., Sanders, E., Tropf, H., 1999. SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line. *Proceedings of Eurospeech'99*, vol. 6, pp. 2699-2702.

Kiessling A., Diehl, F., Fischer, V., Marasek K., 2001. Specification of Databases – Specification of Speakers. Technical report nr D215 available from www.speecon.com.

Marasek, K., Diehl, F., Fischer, V., Kiessling, A., 2001. Specification of Databases – Specification of Corpus and Vocabulary. Technical report nr D213 available from www.speecon.com.

Siemund, R., Höge, H., Kunzmann, S., Marasek K., 2000. SPEECON – Speech Data for Consumer Devices. *Proceedings of LREC 2000.*

Siemund, R., 2000. Functionalities of Speech-Driven Interfaces. Technical report nr D13 available from www.speecon.com.

Van den Heuvel, H., Shammas, S., Moyal, A., 2001. Definition of Validation Criteria. Technical report nr D41 available from www.speecon.com.

Van den Heuvel H., 2000. The Art of Validation. *ELRA Newsletter*, vol. 5(4), pp. 4-6.

Van den Heuvel H., 1999. The SpeechDat-Car Multilingual Speech Databases for In-Car Applications: Some first validation results. *Proceedings of Eurospeech'99*, vol. 5, pp. 2279-2282.

Wells, J. 2002. www.phon.ucl.ac.uk/home/sampa

Winsky, R., 1997. Definition of Corpus, Scripts and Standards for Fixed Networks. SpeechDat II technical report nr SD111 available from www.speechdat.org.