SPEECON - Speech Data for Consumer Devices

Rainer Siemund[1], Harald Höge[2], Siegfried Kunzmann[3], Krzysztof Marasek[4]

[1] Philips Speech Processing, [2] Siemens, [3] IBM, [4] Sony

SPEECON
c/o Harald Höge, Siemens AG, ZT IK 5, Otto-Hahn-Ring 6, D-81739 München, Germany
http://www.speecon.com
info@speecon.com

Abstract

SPEECON, launched in February 2000, is a project focusing on collecting linguistic data for speech recogniser training. Put into action by an industrial consortium, it promotes the development of voice controlled consumer applications such as television sets, video recorders, audio equipment, toys, information kiosks, mobile phones, palmtop computers and car navigation kits. During the lifetime of the project, scheduled to last two years, partners will collect speech data for 18 languages or dialectal zones, including most of the languages spoken in the EU. Attention will also be devoted to research into the environment of the recordings, which are, like the typical surroundings of CE applications, at home, in the office, in public places or in moving vehicles. The following pages will give a brief overview of the workplan for the months to come.

1. Introduction

Telling participants of LREC2000 about the importance of new speech databases is probably as redundant as carrying coals to Newcastle, or, more appropriate, of sending databases to Athens at LREC time. We still dare to do so because a new project, funded within the European Commission's Information Societies Technologies (IST) Programme and put into action by ten industrial partners, has been launched on February 1. SPEECON, a blend of *Spee*ch-driven Interfaces for *Con*sumer Devices, will focus on the creation of speech databases for eighteen languages or dialectal zones. In contrast to the large-scale data collections of the past few years, which chiefly concentrated on telephone material, SPEECON will focus on microphone recordings for the newly emerging field of speech-driven consumer electronics (CE). The various potential application areas are well reflected by the consortium, which is being coordinated by Siemens and further consists of, in alphabetical order, DaimlerChrysler, Ericsson, IBM, Infineon, Lernout & Hauspie, Nokia, Philips Speech Processing, Sony and Temic.

2. The project setup

While in the past speech-driven devices were largely implemented as desktop applications and interactive telephone-based systems such as call centres or directory assistance, the rapid progress in semiconductor technology has now made it possible to further introduce speech driven interfaces into consumer devices. In 1998, applications of this kind generated rather modest revenues of about $17 million world-wide, albeit growing very rapidly at a compound rate of about 54% annually between 1995 and 2005 (Frost&Sullivan 1999).[1] But whereas increasing amounts of data are gradually becoming accessible for 1st generation speech products through agencies like ELRA or projects such as SpeechDat, little has been done in the field of microphone data of the kind needed for the consumer market. Little, that is, both in terms of available data and of methodologies accounting for the potentially wide application areas at home, in the office, in public places or in moving vehicles. SPEECON makes an attempt to cultivate the currently barren field by a number of efforts, largely corresponding to the project's workpackages 1 to 5 and sections 2.1 to 2.5 of this paper. Since most work so far has been dedicated to workpackage 1, the following section will be the most comprehensive one.

2.1 Market analysis

According to a Frost&Sullivan survey, major restraints for a rapidly developing speech command and control market in the past have been, in order of relevance, that

1. accuracy levels in speech recognition were simply not good enough,
2. lack of acceptance of speech recognition in other areas such as desktop dictation and telephony has deflated growth levels in the command and control market,
3. lack of language versions persists,

[1] The figures refer to all embedded command & control applications, divided into consumer electronics proper, automotive, games/entertainment and defense/aerospace markets.

4. people do not want to talk to their PCs, let alone their kitchen appliances,
5. lack of awareness about speech stifles growth in embedded speech markets.

The goal of a project like SPEECON must certainly be to soften the implications of these restraints. The supply of new language material will be discussed in section 2.3, measures to improve recognition rates in 2.4 and the supply of language material in 2.3. The major goal of the market analysis is to design strategies that will create both an increasing general acceptance of voice controled devices and an awareness of the potential speech offers to consumers world-wide. Up till now, the SPEECON partners have devoted most of their time on resolving which kinds of applications are likely to gain market prominence in the near future and which features new voice controled devices are expected to have. Sources used for the market analysis are publicly available market surveys as well as company-internal resources. The idea is to draw a fairly clear picture of potential market segments and to specify which kind of linguistic data will be needed to equip devices with voice control facilities.

The possible areas are tackled in a top-down approach resembling the one used in the various SpeechDat projects (e.g. van Velden, Langmann and Pawlewski 1996, Winsky 1997). First, CE market segments were identified. The areas singled out for the project are:[2]

- mobile phones, mobile hand-held communication devices,
- information kiosks, i.e. rather stationary devices used to access local information, timetables and databases of various kind,
- audio/video devices, both portable or home units to playback, record, receive and transmit acoustic and visual information, including also set-top-boxes and EPG (Electronic Program Guide) control,
- automotive applications, i.e. car mounted devices used to control vehicle functions as well as navigation and car audio,
- toys, mobile and stationary,
- PDAs (Personal Digital Assistants), which may have the full functionality of the PC, but are handheld.

A listing of potential applications in the respective domains follows the account of market segments. An example would be the audio/video branch with applications such as TV, video recorder, DVD, CD-player and recorder, tuner and tape deck. The ensuing step is to define the functionalities these devices are supposed to have, such as, in the case of television:

- audio control (volume, select audio channel), picture control (brightness, colour balance, contrast, colour system, screen format),
- connectivity (channel and input selection), programming (programs, channels),
- attention modes (sleep, stand-by, active, switched off),
- system (language, password protection, help functions),
- time control and
- Electronic Program Guide (EPG), the recent medium for information on film titles, directors, actors, dates and times.

Functionalities of this kind are then translated into broad semantic descriptions such as 'switch the TV on', 'move one channel up' and 'turn the volume off'. Descriptions of this type in turn lead to the definition of the keywords to be sampled during the actual recordings and could read 'activate TV', 'channel up' or simply 'Shsht' for the above functionalities. The process of this first workpackage could be summarised thus:

market segment

application type

functionality

semantic description

command word

The last phase, the definition of the command words for later recordings, however, will be discussed of the following section.

2.2 Specification

A prerequisite for a successful development of spoken language resources is a comprehensive definition of the speech data to be collected. The main issue of this workpackage is to specify databases, which cover adequately the wide range of potential consumer applications. Given a restricted number of utterances recorded from a restricted number of speakers, speech databases have to be tailored from which powerful recognisers can be trained for all envisaged applications.

Main issues of the specification phase are the definition of the corpus, the setup of the recording platforms for the various environmental conditions,

[2] Excluded from the analyses and consequently also from all later stages of the project are areas such as kitchen appliances.

the number of speakers to be recorded, the characterisation of the speakers concerning age, sex and dialect and the definition of transcription criteria. Again, important background knowledge will come from previous SpeechDat projects. As the results to be gathered from the market analysis part of the project are assumed to have an influence on the specification, only four areas have been discussed, though no final decisions have yet been taken, i.e.

1. Recordings will be made according to environment clusters. All target applications deal with either home scenarios such as office, entertainment, children, household areas and living room - or mobile situations such as tourist and public places, car, public transportation.
2. The number of speakers to be recorded will be evenly split between the two scenarios. Since the minimum number of speakers needed to train recognisers adequately are generally considered to be 300, an overall number of 600 speakers, divided between the two recording platforms, will be recorded.
3. The material to be recorded will consist of three types of data, i.e. a) commands spoken in isolation, b) continuously spoken phrases, proper names and digits and c) spontaneous speech.
4. Microphone positions during recordings are supposed to reflect the real distance between speaker and the target applications. Three microphone positions were established, namely,
    1. close, hand-held close to face or head-set
    2. medium: hands-free, desktop (30-100 cm distance)
    3. far positioned (> 2 m)

Table 1 lists the acoustic environments and microphone positions for each application group:

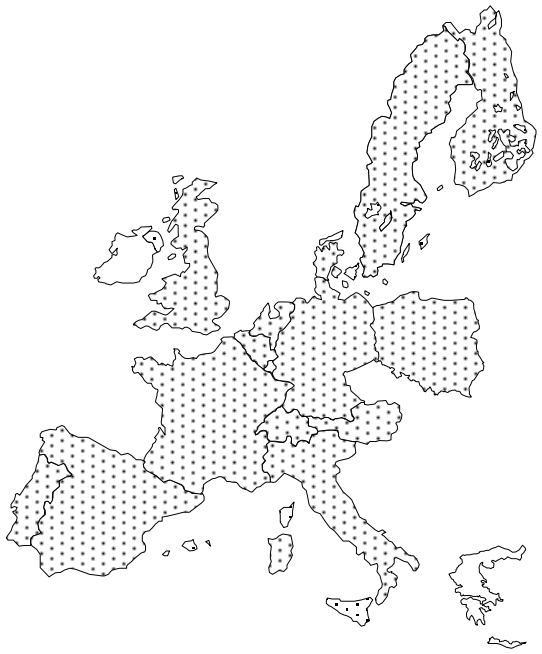
Application	*Acoustic environment*	*Microphone position*
Mobile phone	All	1,2
Info kiosk	Mobile	2
Audio/video	All	2,3
Toys	Children area, living room, entertainment area	2, 3
PDA	All except children area	2

Table 1: Acoustic environments and microphone positions

More details will have been settled by the time of LREC2000.

2.3 Data collection

One of the major objectives of SPEECON is to improve the functionality, usability and general acceptance of CE devices, which will enable users to operate new applications and services in their native tongue. According to the Frost & Sullivan market survey cited above, the lack of available language versions for speech recognition software is among the major constraints which has kept the command and control market from rapidly developing over the past few years (Frost&Sullivan 1999). Since SPEECON is an EU-funded project, the focus of the data collections is clearly on European languages and dialects. Map 1 illustrates the coverage of European language varieties targeted by SPEECON:



Map 1: Coverage of European language varieties in SPEECON

As the map will show, the European languages and dialects to be sampled are Danish, Flemish, Dutch, UK-English, Finnish, French, German, Swiss and Austrian German, Italian, Polish, Russian, Spanish, Swedish and Portuguese. In addition, corpora will be compiled for Chinese, Japanese, Russian, US-English and US-Spanish. The result, a hitherto unknown body of high-quality language material for CE devices, will immediately be used by the SPEECON partners for research on environmental influence on the data, which will be part of the following section.

2.4 Research

Over the past years, developers in the field of speech recognition had basically two choices. Either their system was able to operate with large vocabularies

in rather quiet environments or in noisy surroundings but with only a small set of application words. Examples of both types are desktop and landline telephone applications on the one hand and mobile telephone services on the other. Consumer devices of the type specified in SPEECON, however, need to be operated under virtually any environmental condition irrelevant of the noise level. Mobile phones, for example, will be used both in quiet office environment and in moving cars at 130 km/h. TV sets have to function in living rooms with and without noise emanating from loudspeakers. Information kiosks have to provide the requested details in noisy streets and in train stations with strong hall effects. Further complications arise from the notoriously poor microphone quality of embedded CE applications, small CPU resources and the necessity to operate applications multilingually.

Today's speech recognition systems for consumer devices consequently have only limited capabilities to dynamically adapt to changing acoustic conditions. The main method to minimise degradation of recognition performance currently is to achieve as close a match between acoustic training data and real life scenarios as possible. The sampling of sound data for each potential application area, therefore, is an extremely time-, cost- and labour-intensive undertaking. The possibility to quickly generate training data for target environments of any kind, however, is the key for economic and technological success in a dynamically evolving market.

In order to accumulate knowledge on the acoustic peculiarities of various environments and acoustic conditions, a strong emphasis of SPEECON is on research in this field. Ideally, algorithms will be developed which allow for adaptation of acoustic data between environments. A proper transfer does not only involve background noises of the kind present, for example, in a car vs. a living room, but also has to account for different types of reverberation and features of echo cancellation (Couvreur/Couvreur 2000). The aim of SPEECON's research part is therefore to simulate environments in order to be able to reflect the needs of various devices in the CE domain. The successful transfer of recognition between environments will then be demonstrated by three prototype applications towards the end of the project, to be dealt with in workpackage 5 and the following section.

2.5 Dissemination

The dissemination part of the project deals with making the results of the project available to the public. This will be done in various ways: soon after all data has been collected and research has looked into the acoustic peculiarities of CE environments, the consortium will set up prototypes of three different consumer devices. Each of these devices will have to show two successful transfer steps, either between different languages, i.e. it can be used in, for example, English, Dutch and German, or between environments, i.e. it is operational in one language only but shows high recognition rates in various environments. These prototypes, like overviews of the scope and aims of SPEECON, will be presented at EU meetings, conferences and exhibitions. A few months after the end of SPEECON, all databases will be made publicly accessible through ELRA, while money received from sales will be invested in the compilation of new databases. Throughout the project lifetime, SPEECON activities will be documented through papers like the present one and through the SPEECON website. Documents such as progress reports and publications such as database design, validation criteria, description of room acoustics and noise characteristics as well as other info material will be published as soon as it becomes available. The SPEECON URL is http://www.speecon.com.

3. References

Couvreur, L., and R. Couvreur (2000). On the use of artificial reverberations for ASR in highly reverberant environments. *Paper given at the 2nd IEEE Benelux Signal Processing Symposium, Hilvarenbeek, The Netherlands, March 23-24, 2000*.

Frost&Sullivan (1999). *European Speech Processing Software Markets.* Report #3664-62, p. 5.9.

Van Velden, J., D. Langman, and M. Pawlewski (1996). Specification of speech data collection over mobile Telephone networks. *SpeechDat-II technical report SD 1.1.2/1.2.2*, Appendix III, pp. 14f. Available from http://www.speechdat.org.

Winski, R. (1997). Definition of corpus, scripts and standards for Fixed Networks. *SpeechDat-II technical report SD1.1.1*, Appendix A, pp. 28ff. Available from http://www.speechdat.org.

4. Acknowledgements

SPEECON is funded as a shared-cost project under Human Language Technologies (HLT), which is part of the European Commission's Information Societies Technologies (IST) Programme (Contract number IST-1999-10003).