

## Chapter 8

# Conclusions

The picture sketched in this document seems to be one of lack of standards with the consequent semi-controlled chaos. It cannot be denied that under the present situation considerable resources are spent on tasks like the conversion of corpus formats to internal processing formats. However, even if corpus formats could be strictly standardised, it would still take a very long time before all local software is adapted to that single standard.

On the other hand, the picture shows that there may be good reasons for the use of different designs and different physical representations of speech corpora: the ideal corpus is determined to a very large extent by the application on which one is working. As long as this situation will persist, all attempts to enforce a single standard will be futile. Speech technology has just left the cradle; it is not yet certain how, in what specific direction and along which ways it will develop in the medium and long term. This uncertainty adds to the difficulty to define a single standard for corpora, of which it is as yet impossible to specify their design and contents.

Yet, the picture also shows that, despite the lack of strict standards, general agreement about working standards is developing. As a result it has become increasingly easy to exchange corpora between laboratories. The success of the American LDC has significantly contributed to forming de facto standards. It is expected that the European Language Resource Association ELRA can make similar contributions.

## Chapter 7

# Assessment

The need for standardised procedures for assessing the quality and performance of speech technology is widely recognized. Both the SAM project and NIST have contributed substantially to our insight into the requirements that good assessment procedures should meet. However, it appears to be exceedingly difficult to design assessment procedures that are useful in a wide range of practical situations.

When it comes to designing speech corpora that can support performance assessment essentially the same problem props up that was shortly discussed under Speaker Selection: for a corpus to be suitable to assess technology it should be a representative sample of the speech styles and speech behaviours to be expected in a given application. However, our knowledge about the range of speech behaviours that one should be reasonably able to deal with is extremely patchy and incomplete. As long as this situation persists it will be difficult, if not impossible, to design assessment corpora that allow reliable generalisations from the outcome of a laboratory test to real world performance.

For the moment it is not clear whether it is feasible to build assessment corpora that can be used repeatedly. One condition that should be met is that assessment is carried out in an independent laboratory, that never discloses the test data to the outside world. No such laboratory exists now. The ARPA HLT programme has considered the possibility to set up such an independent assessment laboratory in NIST. However, that proved not to be feasible due to the difficulties in transporting the speech recognition and NLP software from individual laboratories to NIST. As soon as assessment data become public, it is no longer possible to guarantee that the data have not been used during training.

The ARPA HLT programme has shown that it is almost impossible to design several assessment corpora with equivalent levels of difficulty. Until we understand the causes of errors committed by speech recognizers and NLP modules much better than we do now, it will not be possible to compare performance when assessing different recognizers with different test corpora.

## Chapter 6

# Transcription and Validation

The EAGLES Handbook contains an extensive chapter on the topic of transcription and validation of speech corpora. Not surprisingly, much of the contents address corpora designed for long term applications.

As far as short and medium term applications are concerned, there seems to be general agreement at least on transcription standards. Within the SpeechDat consortium transcription standards have been accepted that were derived from guideline proposed by the American HLT programme funded by ARPA.

Rather than repeating the guidelines here, we refer to the Deliverable D3.1.3 on Validation and the internal reports describing the transcription guidelines. It is probably more important here to emphasize that, despite the very large room for adding ever more detail to the transcriptions, there is full agreement on a set of basic requirements. Here too, careful documentation may be as important as anything else. If such documentation is available, data formats are always easily converted between several representation formats.

are necessary for storage. In the speech community, one compression algorithm, **SHORTEN**, designed by Tony Robinson of CUED, has gained substantial popularity, especially in the U.S.A., where the LDC now uses it in most of its newer corpora. The NIST file format and the attendant support software have been adapted to allow for transparent decompressing. Moreover, **SHORTEN** as well as many other 'standard' compression algorithms are fast enough to do decompression on the fly, whenever a file is accessed for processing. This makes compression even more attractive as a means for doubling the size of the data sets which can be kept on line even with low cost equipment. For the SAM file format and attendant support software standard ways of including **SHORTEN** are not yet available. The inherent advantages of 'on the fly' decompression are more than big enough to recommend an adaptation of the support software to allow compression and transparent on the fly decompression also with the SAM file format and software.

Lossless compression using the *Shorten* programme should be considered also for the SAM file format.

Presently, two different file formats seem to prevail for speech corpora, viz. the American NIST format and the European SAM format. For the time being, corpora in the NIST format outnumber those in SAM format. Fortunately, proven software is available to convert between NIST files and SAM files. The major difference between the two formats is in the use of file header. With regard to the storage of speech samples the two formats are identical: speech signals are represented in the form of a byte stream, without additional information interspersed. This form of data storage is compatible with all known speech analysis programmes. It is, however, not compatible with speech file formats that are now being standardised in the multi-media community. Consequently, speech corpora in their present file formats cannot be accessed by the audio programmes that come with multi-media workstations. Up to now, this has not been a major concern, mainly because the way in which speech technology R&D uses speech corpora is different from what happens in multi-media. However, it may well appear that -especially smaller labs, that cannot afford to maintain special software for speech corpora- will request that software be provided with future corpora that converts the internal 'speech technology format' to multi-media file format.

**The issue of conversion of file formats used in speech corpora to file formats used in multi-media applications must be investigated.**

### **5.2.1 File headers**

As said before, the major difference between SAM and NIST format is in the file headers. Essentially, SAM speech files have no headers; the relevant information about the contents of the file is stored in a companion file. The format of these documentation files is simple and straightforward. Moreover, all information in these files is in ASCII format, so that it is easy to read for a human and easy to adapt and correct if the need arises.

NIST files, on the other hand, have the relevant information about the speech contained in a file header. This seems to be the procedure used in older speech signal processing packages (many of which are still in use).

Both approaches have advantages and drawbacks. Probably the most important advantage of the SAM approach is its flexibility and extensibility.

### **5.2.2 One or more tokens in a file**

Many Operating Systems are known to have problems when the number of files in a directory becomes large. One way to counteract this problem would be to store multiple speech tokens (e.g. all tokens produced by one speaker during one recording session) in one file. In order to make individual tokens accessible, an annotation file must then be provided that contains information about the start and end points of the token in the file. This is essentially the approach advocated by the SAM project.

However, virtually all of the corpora produced by the American LDC come with a single utterance per file, and the transcription of the contents of the file in the speech file header. Since the LDC corpora are intensively used as a platform for speech research, also in Europe, and consequently many labs have proven software to support their use of LDC corpora, there appears to be a pragmatic preference for the LDC style of one token per file. Sadly enough from a European point of view, this must lead to the recommendation that, if standardisation is ever to be reached in this area, future corpora should adhere to the LDC format.

**In order to remain compatible with the de facto standard set by corpora produced by the LDC, future speech corpora should opt for storing one token per file.**

### **5.2.3 Compression**

The ever increasing size of speech corpora has raised interest in data compression. Several algorithms are known that can compress speech data in such a way that only roughly half the bytes

## Chapter 5

# Data Base and Storage Issues

A collection of speech recordings can only qualify as a 'corpus' if it is easily accessible for experiments. Therefore, the collection of data should come in a well organised and documented way, and stored (or storable) in a form in which individual tokens are easily accessible.

### 5.1 Data Base design

One obvious way to ease accessibility of data is to store them in a (commercial) Data Base Management System (DBMS). The SAM project has investigated this possibility, focusing attention on Relational DBMSs. Alternatively, speech data might be made accessible using the file handling routines that come with Operating Systems like Unix and VMS. In both cases the conceptual structure in the corpus must be reflected in some way, e.g. in the names of the files or in the organisation of data base records.

Attempts to use RDBMSs for the storage of speech corpora seem to have been outstripped by the growth of the size of the corpora. Now that medium sized corpora require several CD-ROMs for their storage, the use of file manipulation software in the Operating Systems seems to be the only viable approach. However, DBMSs are still in wide use (and will remain so) for storing and accessing information about the status and contents of the files. In practice, this strategy of storing speech signals in files in a directory tree and storing information about the files in data base records is compatible with the decisions made by the American Linguistic Data Consortium.

As the size of speech corpora grows, it will become necessary to include the structure of the directory tree in the overall design scheme, certainly if the CD-ROM must remain compatible with the limitations of the MS-DOS file name conventions. For EUR0M-1 SAM designed a file naming scheme that reserves two character positions to identify the speaker; moreover, it was specified that only the digits 0 to 9 be used to identify speakers. Obviously, this proposal worked well for a relatively small corpus; it does, however, not generalise to corpora like the LDC *Macrophone* or the Dutch *Polyphone*, both comprising over 5,000 speakers.

### 5.2 File formats

File formats for storing speech signals are among the issues that have aroused most debate. Yet, it is not easy to gauge its real importance. It is certainly true that 'standard' processing software assumes 'standard' file formats, but there appears to be little 'standard' processing software around. The practice in most R&D Labs appears to be that they customize their own (often proprietary) speech processing software. As part of that work, they build software to convert several external file formats to their own internal file format. This is feasible as long as corpora come with precise documentation of their file formats.

The procedure sketched above has obvious disadvantages for labs that do not have the resources for customizing and maintaining their own software. For these labs (and the probably form a growing proportion) standardising file formats would be of considerable importance.

surreptitiously recorded speech. The single most obvious advantage of speech recorded in an actual application is its ecological validity.

So called *Wizard of Oz* designs are the experimental paradigm that comes closest to recording in real applications. When speech is recorded in this way, it is probably possible to debrief all speakers, and ask for consent to publish the recordings. One should be aware, however, that in most *Wizard of Oz* experiments speakers know that they are subjects in an experiment. Moreover, these subjects follow scenarios that they might never have dreamed up themselves. This may have a considerable impact on their speech behaviour. Last but not least, details of the speech behaviour may also be influenced by the well known experimenter effect: the experimenter may induce specific behaviors, e.g. by the way in which instructions are formulated (and which may suggest specific ways for expressing times and/or dates, to give just one example).

## 4.5 Speaker selection

Speaker selection is one of the most important issues in corpus design and collection. Unfortunately, no general guidelines can as yet been given to support future corpus designers. Perhaps, speaker selection is the aspect where the impact of the goal for which the corpus is collected weighs most heavily.

In methodological terms, the single most important issue in speaker selection is that the eventual set of speakers must be representative for the population towards which a research or development project is directed. That need not be the general public; for instance, in military applications the population characteristics may differ substantially from civil applications in the PSTN.

The problem of defining the population and of the sample to represent it is aggravated by the present lack of knowledge about the importance and the composition of the personal, physical, social and psychological dimensions that span the speaker space. In other words: as long as we do not know how to describe the population of interest in the terms that really matter (*viz.* the terms that allow us to predict how speech behaviour will vary) it will remain difficult to be confident about speaker selection criteria and guidelines.

The EAGLES Handbook contains an extensive description of speaker characteristics that might have a non-negligible impact on speech behaviour. The most important among them are certainly sex, regional dialectal background, and level of education. For some applications age and physical health may also come in as significant factors.

Although one might think that a speech recognizer trained on speech of as many speakers as possible should outperform a recognizer trained on the same amount of speech, but recorded from a small number of speakers, BBN has reported contradictory results: for the Wall Street Journal corpus they find that training on 12 speakers is as good as training on 84 speakers, as long as the total amount of training material is constant. They also report that combining the training material of the 12 speaker and 84 speaker sets improved recognition performance only marginally. This result suggests that there are limits to the amount of information that our present recognizer training algorithms can extract from the training material. Alternatively, one must conclude that the boundary between 'few' and 'many' speakers is at a number (substantially) above 84.

Informal reports of speech technology engineers support the observation that the claim 'the more speakers the better' may not always be true.

When collecting corpora for *speaker* instead of *speech* recognition the speaker selection issue is completely different. However, at the present state of our knowledge it is not possible to give strict guideline for this application either. That is to say, not beyond the general, but essential remark that speakers should be recorded repeatedly, over a time interval of at least three months, and under different recording conditions.

A speech style that is closely related to read speech, and that can be used to advantage in recording corpora consisting of isolated words and simple connected speech, is to ask subjects to repeat the words and/or expressions spoken to them by the recording workstation.

#### 4.4.1.1 Prompting techniques

When the corpus designer wants to record specific words or expressions, the intuitively simplest way is to ask speakers to read these words or expressions. Using read speech is certainly easy if the speakers are requested to come to a recording studio, or if the experimenter pays personal visits to the speakers. In fact, this was the way in which most older corpora, like EUROM-0 and EUROM-1, have been collected. Also, when using read speech one can monitor the speech production during recording in detail and ask the speaker to repeat each utterance that contained errors or disfluencies. It has been argued that read speech may lead to unnatural productions. However, we know of no failures in developing speech technology applications that could clearly be attributed to this problem.

When all speakers can be contacted personally, it is advisable that the text to be read is displayed on a computer screen, instead of printed on paper. Electronic displays are much more flexible than printing on paper. For instance, semi-automatic re-prompting of all items flagged as questionable or erroneous is only feasible with computer controlled prompting. When a quiet terminal is used, computer prompting can also contribute to a better S/N ratio, because it prevents paper rustling.

However, in recording corpora comprising some 5000 callers on the telephone personal contact between the experimenter and all speakers is impossible. In these circumstances there is hardly an alternative for sending prompt sheets to all prospective callers. This situation is not likely to occur when collecting corpora for dictation R&D; it is, however, the rule rather than the exception in collecting telephone speech corpora.

Using read prompts it is possible to collect not only isolated words and connected words, but also continuous speech.

When collecting a precisely defined set of words, the only alternative for read speech is asking the speakers to repeat utterances spoken to them by the live experimenter or the recording platform. In this way isolated words and probably also connected words can be recorded. It is questionable whether this prompting technique lends itself to recording continuous speech. Experience gained in psycholinguistic research suggest that many people have great difficulty in repeating even medium length sentences verbatim.

It has been argued that prompting words by speaking them to the subjects is at least as objectionable as asking them to read lists of items. However, many speech technology engineers report good results with corpora collected with this prompting technique. When recording speech in running cars, repeating words spoken by the experimenter may be the only way of prompting that is compatible with traffic safety considerations, certainly when the speaker must be the driver.

It has been attempted to prompt words by asking the speakers to answer questions. However, in order to be sure that the exact word intended is indeed spoken, the questions must be of a debilitating simplicity. And even then there can be misunderstandings, leading to deviating responses.

#### 4.4.2 Spontaneous speech

Non-read speech can, at least in principle, be recorded under a large number of different circumstances. Up to now, few corpora containing spontaneous speech are available. By far the largest and best known is the ATIS corpus; it can be argued, however, that the speech in this corpus is much better prepared than speech produced in normal inter-personal communication. Such an objection can certainly not be made against the American English *Switchboard* corpus, which contains only spontaneous conversations, be it between strangers and recorded over the telephone.

One obvious way of recording unprepared speech would be to store the speech that clients produce in interaction with some automated service. Several examples of (proprietary) corpora collected in this manner are under development. A major drawback of this recording technique is the limitation on distribution, due to privacy laws that prevent the recording party from publishing

Keeping items embedded in audible background noise does make corpus transcription and validation more time consuming. However, careful monitoring of extraneous noise during corpus recording, and repeating distorted items until a clean token is obtained, is likely to be at least as time consuming and expensive. 'Realistic' corpora, then, have the advantage that they can also be used to train and test noise and garbage models.

## 4.2 Connected words

Some applications are very tedious if the speaker can only use isolated words. One such application is entry of card or phone numbers. For these tasks it would be preferable if the speaker would be allowed to speak the numbers without intervening pauses. This type of speech is known under the term *connected* speech.

Several corpora exist that contain connected speech, mainly connected digits. Another type of connected speech that is quickly gaining popularity comprises letter names, to allow spelled input. In designing connected speech corpora, care must be taken to include all frequently used expressions, which may differ between languages or between cultures. Expressions that must be taken into account include *double*, *triple*, etc. In some cultures reading (or speaking) phone numbers is customarily done in terms of numbers between 0 and 99, instead of in terms of the digits 0 to 9.

## 4.3 Continuous speech

Continuous speech corpora have been in use by the dictation R&D community for quite some time. Continuous speech recordings for telephone applications are from a much more recent period.

In the dictation community a number of very large corpora of -mostly read- speech have been recorded. Most of these corpora pertain to American English. However, there is a British English version of the Wall Street Journal corpus. Newspaper text corpora Large corpora of read speech might also be used for deriving c.q. optimizing rules for speech synthesis, especially if very large amounts of speech were available from one speaker.

One of the largest corpora containing continuous telephone speech is the POLYPHONE corpus, which is available for American English (under the name MACROPHONE) and for Dutch. Corpus collection for other languages is under way, e.g. in the SpeechDat project.

## 4.4 Speech styles

A person's speech behaviour spans a wide range, under the influence of a set of context factors that is still ill understood. It is quite likely that different subjects will react differently to the same situational factors. For instance, people who are used to reading text out loud will have less difficulty in reading words, sentences or texts than persons who practically never read, let alone read texts out to other persons.

Read speech is certainly not the only speech style, and for most persons it is certainly not the most common style. However, it is easy to see that 'spontaneous' speech, as opposed to read speech, is a very vague and ill-defined concept. 'Spontaneous' speech defined as 'non-read' speech may vary from formal and short answers to factual questions to conversational chat. For the time being, scientists have not succeeded in coming up with a useful taxonomy of non-read speech.

### 4.4.1 Read speech

For a number of goals and applications corpora consisting of read speech are probably ideal. Read speech is probably the fastest way to collect tokens of small sets of words, spoken by a large number of speakers. Recording read speech is certainly the only feasible route towards large amounts of tightly controlled speech.

## Chapter 4

# Linguistic contents

The linguistic material in a speech corpus can -and actually does- vary widely, with a small number of words spoken in isolation on the one extreme to extemporaneous conversational speech on the other. Up to now, most corpora focus on isolated words, with some corpora also containing connected words and continuous speech.

Quite another aspect of linguistic content regards the distinction between speech styles. Here, read speech should be distinguished from more spontaneous types of speech production.

A third issue, somewhat more remotely related to linguistics, is the selection of speakers to be recorded.

The linguistic content of a corpus is very strongly determined by the application for which the corpus was collected. Due to this dependence, it is not possible to give general guidelines for determining 'standard' or 'optimal' linguistic contents of a corpus.

In terms of corpus content there has long been a big difference between two generic types of applications, viz. speaker independent speech recognition over the telephone (focusing on small vocabularies of isolated words) and office dictation (focusing on large vocabularies of words spoken in isolation or continuous speech). Although the gap between these application types is narrowing, it will persist during the rest of this century, if only because of the problems with signal quality in the cellular telephone networks.

### 4.1 Isolated words

Many corpora contain small sets of isolated words. One very popular set is comprised of the digits 0, 1, ..., 9. Another popular set includes the words for *yes* and *no*. Yet other sets comprise frequently used command words, like *start*, *stop*, *add*, etc.

Experience in designing applications has shown that, except for the digits and *yes*, *no* expressions, command words are often highly application specific. This makes it very difficult, if not impossible, to collect standard corpora that are guaranteed to contain all command words relevant for a new application.

Experience with applications directed at the general public has shown that especially occasional users of a service relying on speech recognition have great difficulty in using isolated words. More often than not they add other expressions to the word, like hesitation sounds, or they say *yes*, *please* instead of *yes*. Older corpora were carefully cleaned from extraneous speech and background noise. That (time consuming) practice may be advantageous as long as one must train optimal models for the words the recognizer is supposed to handle, but it turns into a disadvantage as soon as it comes to training models for extraneous noise. Also, cleaned-up corpora are not very useful for assessing the performance of a recognizer in such a way that it is safe to make generalizations towards performance under realistic operation conditions.

**It is strongly recommended to keep errors, hesitations, extraneous speech and other background noise in the corpus.**

application would be a transport information service that is called by a person who is desperately seeking ways to make it in time to the airport (or any other location). To our knowledge, up to now no corpora of speech under emotional stress have been recorded in the EU.

### 3.2.1 Car recordings

Recordings in (running) cars can be made with two different goals in mind:

- development of speech recognition in a local terminal
- development of speech recognition in a central switch.

These goals require different recording set-ups. If the major purpose of corpus collection is the training of hands free command/control in the car, it is advisable to make wide band recordings, using a recording workstation in the car. If the major purpose is speech recognition in a (central) network switch, one should record in a switch, in order to be able to make good estimates of the impact of bit errors in transmission (in digital cellular networks) or fading (in analogue networks). One might even consider dual recordings, i.e., wide band in the car and simultaneous recording of the telephone bandwidth channel in (or at the far end of) a switch.

Environmental noise conditions are very likely to change during recording sessions in a car. It is strongly recommended to continuously keep track of the changing conditions during recording sessions.

One factor which is especially important when making car recordings -and therefore should always be carefully monitored and recorded- is the status of the windows (and the sun roof, if present). Environmental noise conditions change dramatically when windows are opened or closed.

Another factor which has a major impact on the background noise is the speed at which the car is running. Often, this correlates with the kind of traffic the car is in. It is advisable to make recordings in at least four traffic conditions, i.e.,

- (1) parking with engine running stationary,
- (2) running at slow speed in town centre,
- (3) running at a speed of 80 km/h on major routes,
- (4) running at a speed of 120 km/h on a highway.

Each of these four traffic conditions can be combined with open/closed window condition.

When making recordings in a running car, care must be taken not to distract the driver's attention to such an extent that safety may be compromised. The best way to reach that goal is to have the speaker sit in the front passenger seat. However, this situation may not be valid in the case of hands free car phones, because the microphone is bound to be fixed near the driver's seat. For hand held car phones there is no objection against the speaker being in the passenger seat.

It may be argued that recordings made in a switch from a speaker in a (running) car should always be accompanied by a wide band recording of the speech in the car, or by a record of the field strength of the base station and the hand-over signals. Only when these additional data are available will it be possible to distinguish between the adverse effects of environmental noise and the (perhaps even worse) effects of transmission errors.

**The recording conditions must be carefully described.**

If recordings are made in a fixed (studio-)room, it is worthwhile to document room acoustics by recording the room's impulse response.

If recordings are made under conditions that may vary during one session, additional means must be provided to log the changing conditions during the session.

### 3.2.2 Recordings under extreme physical conditions

For military applications corpora of speech recorded from speakers who are experiencing extreme physical or emotional stress are necessary. Physical stress is caused by *g*-forces exerted on pilots in fighter jets at high accelerations. Alternatively, physical stress may be due to extremely high background noise levels, which do not only occur in jets, but also in helicopters. Speech corpora for military developments are not likely to be made publicly available; therefore, we will not go into recommendations for the collection of such corpora.

Speech under high emotional stress may also be of interest for civil applications. One such

where the direct sound field has maximum power.

For telephone applications the microphone is mounted in the handset. Callers can vary the distance and angle between lips and microphone to a considerable extent.

The EAGLES Handbook gives extensive descriptions of a large number of microphone types. Therefore, we will not go into further details here.

In addition to the various types of microphones in telephone handsets (mainly carbon button in the older corpora, with an increasing share of electret microphones in the newer corpora) close talking magneto-dynamic microphones have been very popular for recording speech corpora (e.g., Wall Street Journal, ATIS, etc.). Part of the Wall Street Journal corpus is simultaneously recorded with two microphones, i.e., with a flat mounted table top microphone along with the close talking microphone.

Several experiments have been conducted to establish the optimal placement of the microphone for hands free telephones in cars. As far as we can see, no agreement has been reached about optimal placement. Moreover, it is not obvious that optimal placement is independent of the type and make of the car.

One issue that has come up a number of times is the use of microphone arrays as a means to improve S/N ratio in speech recordings. Up to now, no speech corpora with multiple parallel microphone channels are available.

No general guidelines can be given for the use of microphones in recording speech corpora.

The choice of the microphone, the mounting of microphone, distance from lips to microphone, etc. must be carefully documented, for each recording session in a corpus.

### 3.2 Environmental conditions

With respect to the choice of environmental conditions for speech corpus recording two schools must be distinguished:

a group of scientists who have their major background in (psycho-)acoustics prefer to make completely clean recordings in anechoic rooms; room acoustic conditions, background noise, and transmission distortions are then added by means of linear and non-linear signal processing as needs arise.

scientists with a speech technology background stress the importance of making recordings under conditions which are as close to the application conditions as possible.

Both approaches have their strong and weak points. The major shortcoming of anechoic recordings is that they do not contain the effects on the speaker's behaviour due to the environmental conditions. It is well known that talkers raise their voice level in noisy conditions. Other conditions may have different effects on the talker's behaviour, but there is little or no documentation. Some believe that the unusual circumstances prevailing in an anechoic room have themselves strong effects on speaker behaviour. It has been said before that the higher order statistics of temporal variability of background acoustics and channel distortions are very difficult, if not impossible to simulate.

The weak point of real-life recordings is that they may not as valid and interesting for other applications than the one for which they were obtained. Specifically, recordings made in noisy environments cannot easily be used for fundamental phonetics research. Also, recordings in the cellular telephone network may not be useful for applications in the fixed network, due to the irremovable effects from bit-errors during transmission.

There are two recording conditions that require special attention, viz. recordings in cars and recordings under extreme physical (and/or emotional) stress.

## Chapter 3

# Physical conditions

In this chapter an overview is given of the physical conditions under which speech corpora have been recorded in the past and will be recorded in the foreseeable future. By doing so, due attention will be given to the influence that these physical conditions may have on the use of corpora for the development of several types of applications.

For this document *physical conditions* are taken as comprising both the transducer (microphone) and the acoustic quality of the room in which the recordings are made. Also, attention will be paid to transmission conditions in telecommunication networks.

### 3.1 Microphone

In this document only acoustic speech signals will be considered. It is acknowledged that some corpora comprise additional signals, like the *laryngograph*  $L_x$  signal, because virtually all short and medium term applications of speech technology are limited to the use of the acoustic speech signal. The only exceptions that we can see are very specific aids for the disabled (for the development of which narrowly focused corpora are required) and military applications (e.g. the use of contact microphones in oxygen masks); if corpora for military applications do exist, it is not likely that they will be made publicly available.

Microphones can be classified along a number of dimensions, e.g.

1. according to the physical mechanism used to transform air pressure into electrical voltage  
Along this dimension one can distinguish
  - condenser
  - electret
  - magneto-dynamic
  - carbon button
2. the directional sensitivity  
A number of different directional characteristics have been implemented. For the collection of speech corpora the major distinction is probably the one between microphones that are sensitive for signals from all directions versus mono-directional microphones.
3. the packaging and mounting of the transducer  
The distance and angle between the speaker's lips and the microphone are a very important factor determining the S/N ratio of speech recordings. If the microphone is mounted independently of the speaker, this distance and angle may vary. Lightweight microphones can be attached to a headset to be worn by the speaker, ensuring that the microphone distance is kept relatively fixed. Another aspect of transducer packaging which is important for head mounted microphones is the suppression of turbulence due to DC flow. Close talking microphones do avoid such turbulence, so that they can be used in a position very close to the lips,

2. from a (digital) line in the fixed network. In this situation a transcoding of GSM to A-law is necessary. The obvious advantage is that use can be made of any recording workstation suitable for making recordings off the telephone network.

Up to now, the practice seems to be that GSM signals are recorded after transcoding to A-law (or after decoding to analogue) signals.

**Best Practice guidelines for recording GSM signals should be designed.**

made to the recording platform.

It is recommended NOT to use CODECs.

### 2.3.2 Digital source signals

Up to now, digital source signals have mainly been limited to recordings off a digital telephone connection. In the fixed landline network one of two coding schemes is used: u-law (in the U.S.) and A-law (in Europe). Both coding schemes use 8 kHz sampling and logarithmic companding of amplitude values to fit in 8 bit words. Thus, the bit rate is 64.000 bits/s.

Before companded signals can be used for processing they must be converted to sample values on a linear amplitude scale. The conversion is simply effected by table look-up.

u-law signals recorded in the American PSTN contain signalling information in addition to speech information. This is not the case with A-law signals recorded in the European PSTNs.

Several recording workstations exist that can be hooked up to the PSTN. In doing so, software must be provided for signalling: detecting that there is a call, establishing the connection, detecting a break of the connection by the caller, freeing the line.

Signalling software must be adapted to the signalling protocol used in the network. Unfortunately, these protocols may differ between the countries, even in the EU. Where available, the Euro-ISDN protocol should be used, so that maximal transparency of the software of the recording platform can be obtained.

There are a number of ISDN boards that come with the necessary drivers for PC's under OS/2 and Windows. Many modern Unix workstations come standard with an ISDN port. However, up to now no single workstation manufacturer is known who provides driver software that support the ISDN speech transmission protocol. All drivers that are available only support data transmission protocols. Software support for ISDN speech transmission is available for SUN workstations from an independent Swiss software house.

#### 2.3.2.1 GSM

The GSM network is without doubt the fastest growing cellular network in the EU. In comparison with the fixed network, GSM poses two additional problem:

1. compression

GSM is based on a CELP coding algorithm that allows to obtain reasonable perceptual signal quality at a rate of 13 kbits/s<sup>1</sup>. However, as has been mentioned before, such coding effects irrecoverable distortions. In the case of GSM these distortions are certainly audible. It is not yet known how they affect feature extraction in systems for automatic speech and speaker recognition.

2. fading

Due to varying signal levels at the base station, GSM signals can suffer from bit-errors. Complete packets may disappear, because they are mutilated beyond repairability. It appears that it is extremely difficult to obtain good estimates of the higher order time statistics of bit errors. This makes realistic simulation of transmission problems not feasible.

There are two ways for recording GSM signals:

1. in the GSM network, resulting in a recording of the original GSM codes. No workstation is known that supports this type of recording.

---

<sup>1</sup>The European GSM 06.10 provisional standard for full-rate speech transcoding, prI-ETS 300 036, which uses RPE/LTP (residual pulse excitation/long term prediction) coding at 13 kbit/sec. GSM 06.10 compresses frames of 160 13-bit samples (8 kHz sampling rate, i.e. a frame rate of 50 Hz) into 260 bits.

For wide band applications a sampling frequency of at least 16 kHz should be used.

### 2.3.1.2 Bits/sample

Popular choices for bits/sample have been 12, 14, and 16. There may still be some very old digitized recordings that have only 8 bits/sample.

The major issue connected to the word length is *dynamic range*: the more bits per word, the larger the range between the loudest and softest passages that can be accommodated without peak clipping in the loud passages and without the weakest passages drowning in background noise. Although the dynamic range in normal speech is around 35 dB, a range that can be covered adequately by 12 bits, the use of wider words is recommended, due to considerable differences in average loudness between speakers, between recording sessions and between microphone amplifiers. Using 16 bits instead of 12 allows one to obtain reasonable signal/noise ratios for a wide range of recordings, without the need for careful setting and monitoring of input gain or input attenuation. Moreover, in computers with 8 bit bytes the storage needed for 12 bit words and 16 bit words is equivalent, unless one decides to use a storage scheme that packs 4 12 bit samples into 3 16 bit words.

In a number of telephone speech corpora recordings have been made off an analogue telephone line. Despite the fact that the dynamic range of telephone connections is small, often 16 bit samples have been used. This makes sense for exactly the same reasons as mentioned above.

When the A/D board offers a choice for bits/sample, it is recommended to select the largest number available.

### 2.3.1.3 Filtering and coding

Most A/D boards do more than just analogue to digital conversion. Almost all boards come with low pass filters, the cut-off frequencies of which can be adapted to the sampling frequencies. Some boards come with many additional signal conditioning features, like gain control and high pass filtering.

Especially for the cheaper boards it is necessary to check their performance. It has been observed that the actual signal/noise ratio obtained is substantially lower than what one might expect from the number of bits/sample. This is due to electronic noise picked up by the analogue part of the interface. Checks can be made by recording a file with short circuited input, or with the microphone connected but switched off. The number of bits under the noise floor can then be established by making simple statistics for the sample values. Alternatively, a pure sine wave can be recorded (if the board comes with a line input). S/N ratio can then be determined from a spectral analysis of the input.

Some A/D boards contain coder-decoder (CODEC) hardware, that effects a non-linear conversion from analogue voltage values to digital sample values. The use of a CODEC has the same effect as the compression algorithms in DCC and mini disk players: the amount of disk space needed to store the signals is diminished, at the cost of irrecoverable -be it sometimes inaudible- distortions of the signal.

In using additional features of the A/D board care must be taken not to unnecessarily distort the input signal. Any use of such features must be carefully documented.

Unless there are good reasons for doing otherwise, it is recommended that no signal conditioning features be used.

Fixed checking procedures should be followed to establish the real S/N ratio during recording. This procedure must be repeated every time a change has been

harmful.

Similar remarks seem to apply to recordings off the telephone network. Several different telephone interface boards, supported by different software packages are in use in Europe and elsewhere.

Although the speech signal is always produced as an analogue signal, it may still appear as either an analogue or a digital signal at the input of the recording workstation. This is especially true when recording speech signals from the Public Switched Telephone Network. Therefore, it makes sense to distinguish between recordings of analogue and digital source signals.

### **2.3.1 Analogue source signal**

The situation where the source signal is analogue has been by far the most common one. In terms of applications it applies to all situations where the microphone is directly connected to the speech processing device. Applications include recognition in the terminal connected to the telephone network, but also to dictation, command and control tasks, etc.

If the source signal is analogue, an A/D converter is needed to interface the analogue speech signal to the digital computer world. Many plug-in boards with A/D converters are available. In the past, these boards tended to be relatively expensive and difficult to use. In that situation it was of considerable importance that SAM made attempts to provide standardised software for a board that would satisfy most needs. Now that the price of the boards has dropped dramatically and most boards come with proven software that supports applications like recording speech to disk, the need for common standardised software is no longer urgent.

#### **2.3.1.1 Sampling frequency**

In the past, several different sampling frequencies have been advertised as 'standard'. However, the practice is that a range of sampling frequencies has been used. Attempts to set a fixed standard have failed, probably because the bandwidth requirements of different applications differ.

Sampling frequencies that have been used in the past include

8 kHz, mainly in work directed towards telephone applications

10 kHz, a sampling frequency that has long been popular in academic speech research; it covers the frequency band up to 5 kHz, known to be the band containing all information about voiced sounds and most information about unvoiced sounds

16 kHz, the sampling rate that gained popularity when the costs of storage began to decrease rapidly; compared to the 10 kHz rate this choice has the advantage that it covers a band wide enough to contain essentially all speech information

20 kHz; here the same considerations apply as for 16 kHz, but in a world in which 10 kHz was the standard, choosing 20 kHz facilitates sampling rate conversion towards the old standard

40 kHz+, the sampling rates used in DAT and CD recordings

Given the wide availability of powerful computers, the choice of sampling frequency has lost some of its importance, because rate conversion can be effected with negligible loss of recording quality (although one should carefully check the performance of a given rate conversion programme). However, a couple of things must be kept in mind:

raising sampling frequency after digital recording does not restore the frequency band above the cut off frequency in the original digital signal

rate conversion to exactly 10 kHz may be quite complicated if the original rate is e.g. 44.4 kHz.

One should be aware that not all A/D boards allow a very flexible choice of sampling frequencies.

by Sony) that reduce storage by a factor of 4. This reduction is obtained by means of clever coding strategies, that take the limitations of the human auditory perception mechanism into account. Although experiments have shown that compressed and uncompressed recordings cannot be distinguished auditorily, the effect of the compression algorithms on the output of the signal processing algorithms that are customary in speech science and technology has not yet been investigated. Depending on the type of processing, such effects may or may not exist. Moreover, the compression algorithms used by Philips and Sony differ.

If recordings are made onto tape, the use of industry standard DAT is recommended.

Until more insight is obtained into the effects of the compression algorithms, the Philips DCC and Sony Mini Disk should not be used to record speech corpora.

If recording onto direct access disk is feasible, DAT is NOT the medium of preference.

### 2.3 Digital recording onto direct access disk

Since the advent of large, fast and cheap direct access disks, they have become the preferred medium for recording speech corpora. Direct access disks share all the advantages of digital tape recordings, but lack their limitations. The major remaining limitation of digital disks is still their storage capacity, but even this limitation is losing much of its practical meaning: recently disks with storage capacities of 4 Gbyte and 9 Gbyte have been introduced. However, one must be aware that there are still large price differences between media costs, with tapes obviously having the advantage.

The availability of very large capacity direct access disks suggests the following working procedure:

record the raw speech signals onto direct access disk, while at the same time building an index of items recorded

select, validate and transcribe the items to be kept for later R&D work; if necessary, these items -with the attendant index- can be copied to another disk

after the first processing stage, the original recordings can be copied to a cheap -and therefore probably slow- back-up medium, and the original recordings can be deleted, thus freeing the disk for making new recordings

for exchange of corpus material CD-ROMs or DAT can be used

In the last decade, a number of recording workstations have been built. With few exceptions -which use Unix workstations- these recording stations are all based on IBM compatible PC's. Some recording platforms which are especially built for recording telephone signals require OS/2 as Operating System.

In practice, many different software/hardware combinations are in use worldwide to support the recording of speech corpora. One widely used package is known as SAMPEC, one of the results of the ESPRIT Project SAM. Compatible with the state-of-the-art in the late eighties, SAMPEC is based on an affordable PC, but it also includes a relatively expensive DSP board. However, the software architecture allows to replace that board by modern, low-cost sound boards.

Recently, the advent of hardware/software platforms to support multi-media applications has increased the number of possible solutions considerably. It is unlikely that a common standard will emerge in the foreseeable future. But as long as it can be guaranteed that labs do not waste large amounts of time and effort in unnecessarily building custom solutions, and as long as the recorded files are compatible with common standards, this lack of standard recording platforms may not be

## Chapter 2

# Physical Recordings

The oldest speech corpora have been recorded on analogue magnetic tape. The analogue recordings were subsequently digitized and stored in computer readable form. Since the middle of the eighties, direct digital recording, either on tape or on direct access disks became feasible. Presently, digital recording onto direct access disk is the standard.

### 2.1 Analogue tape recordings

Until the middle of the eighties there was no affordable hardware that would allow recordings of many minutes worth of speech directly onto digital media, regardless of whether that was sequential access tape or direct access disk. Thus, there was no alternative for recording speech onto analogue magnetic tape, that was cheap and easy to handle, in a way. After completion of the recordings, the parts (e.g., words, utterances) that were selected for further research were individually digitized and stored onto computer readable tape or disk.

It is well known that analogue tape has considerable drawbacks. The signal/noise ratio of the recordings degrades over time. Perhaps more importantly, editing analogue tapes to select the items to be digitized and digitizing short stretches of speech is a very time consuming (and probably also error prone) process. Therefore, it is not surprising that analogue tape recordings were abandoned as soon as large scale digital recording became feasible.

Now that affordable digital recording is widely available, analogue tape recording should be absolutely avoided.

### 2.2 Digital tape recordings

The first digital audio recorders used video tapes as their medium. In fact, the oldest digital tape recorders were video cassette recorders equipped with an A/D and D/A conversion box. These video machines have been superseded by Digital Audio Tape (DAT), which -contrary to the older video systems- adheres to an industry standard. This makes DAT tapes a much better medium for data exchange than the older video cassettes.

Digital tapes are relatively immune to deterioration of signal/noise over time. However, they share one disadvantage with analogue tapes, viz. the fact that they are essentially sequential media. Searching for a passage on a DAT can take a long time. Moreover, there are essentially no means for storing a useful index of the recorded material on the tape itself. This requires that a separate index be maintained.

All digital tape media have in common that they use high sampling frequencies (40 kHz+) and many (typically 16) bits per sample. Consequently, speech signals on digital tape consume very large amounts of storage space. In many respects, the storage is not warranted by essential information in the signals.

Recently Philips and Sony have introduced digital recording media (tapes by Philips, mini disks

# Chapter 1

## Introduction

This Deliverable is, in a way, a summary of part of the work done by the *Spoken Language Working Group* in the EAGLES project. However, it is equally valid to say that scientists working in SpeechDat made essential contributions to the EAGLES Handbook.

This document is best read with constant reference to EAGLES Handbook. The latter document contains much more and detailed information, if only because its size is an order of magnitude larger. Although for its contents this document draws liberally from the EAGLES Handbook, this text was written completely anew, in order to maximize homogeneity of style and contents.

The organisation of this deliverable differs somewhat from the chapter- and section structure of the EAGLES Handbook, mainly because this deliverable does only deal speech corpora collected for short and medium term applications. The EAGLES Handbook, on the other hand, also covers corpora that should support long term developments. In its turn, this document addresses a number of issues that were only mentioned in passing (or not at all) in the EAGLES Handbook –because they seemed to be too closely linked to specific types of applications – in somewhat more detail. Although we have made an attempt to balance the attention for the generic application types, we may not have succeeded in avoiding a bias towards telephone network applications.

This deliverable adheres to the style of the EAGLES Handbook in that it attempts to give recommendations wherever that is possible.

**Recommendations are printed in typewriter font.**

DEL 3.1.1.1

REPORT ON WORKING STANDARDS FOR SPEECH DATA  
BASES DIRECTED TOWARDS SHORT AND MEDIUM TERM  
APPLICATIONS

Lou Boves\*

Els den Os<sup>+</sup>

\*SPEX, Nijmegen University, KPN Research <sup>+</sup>SPEX

October 15, 1995