INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 4, 127–143, 2001

Annotation in the SpeechDat Projects

HENK VAN DEN HEUVEL AND LOUIS BOVES
SPEX, A2RT, University of Nijmegen, The Netherlands

ASUNCION MORENO
UPC, Barcelona, Spain

MAURIZIO OMOLOGO
IRST, Trento, Italy

GAËL RICHARD
Philips Consumer Communications, Montrouge, France

ERIC SANDERS
SPEX, A2RT, University of Nijmegen, The Netherlands

Received July 28, 2000; Revised March 12, 2001

Abstract. A large set of spoken language resources (SLR) for various European languages is being compiled in several SpeechDat projects with the aim to train and test speech recognizers for voice driven services, mainly over telephone lines. This paper is focused on the annotation conventions applied for the Speechdat SLR. These SLR contain typical examples of short monologue speech utterances with simple orthographic transcriptions in a hierarchically simple annotation structure. The annotation conventions and their underlying principles are described and compared to approaches used for related SLR. The synchronization of the orthographic transcriptions with the corresponding speech files is addressed, and the impact of the selected approach for capturing specific phonological and phonetic phenomena is discussed. In the SpeechDat projects a number of tools have been developed to carry out the transcription of the speech. In this paper, a short description of these tools and their properties is provided. For all SpeechDat projects, an internal validity check of the databases and their annotations is carried out. The procedure of this validation campaign, the performed evaluations, and some of the results are presented.

Keywords: annotation, transcription, SpeechDat, annotation tools, validation

1. Introduction

In order to test and especially train automatic speech recognition (ASR) systems very large amounts of speech data are generally required. Such spoken language resources (SLR) have been produced in ever increasing numbers during the past decade. In the SpeechDat projects a great number of SLR have been produced for a large variety of European languages.

The purpose for which the SpeechDat corpora have been built and the dominant ideas about the best way to develop ASR for telephony applications in the early nineties—the time when the plans to create the corpora took solid form—have, by necessity, had a decisive impact on the design and the annotation of the corpora.

This paper aims at evaluating the approach used for the annotation of the SpeechDat SLR family.

SpeechDat, like a number of companion projects in the USA, Japan, Korea, and China, had its origin in the need for corpora that can support the development of flexible vocabulary, speaker independent ASR over the telephone. In the early nineties this meant 'robust' recognition of digits and digit strings, money amounts and application-specific words and expressions. The ASR device was seen mainly as a replacement of the DTMF (Dual Tone Multi Frequency) detector in a fully system-driven interaction between a caller/customer and an advanced IVR (Interactive Voice Response) System. This is reflected in the design as well as the recording procedures used to collect the predecessors of SpeechDat (the American-English Polyphone corpus Macrophone (Bernstein, et al., 1994) and the Dutch Polyphone corpus (Den Os et al., 1995)). The design and recording procedures for the SpeechDat corpora are similar. Most of the speech is read aloud from a prompting sheet sent to the speakers before the recording session. The utterances are relatively short; some consist of just a single word, while the longest are typically sentences consisting of 10 to 12 words. The short utterances were designed for training and testing isolated word recognition systems (requiring enough tokens of each word to obtain training and test sets of reasonable size). The sentences were mainly meant for training sub-word models, that could then be tested with the short utterances (and also with part of the sentences, of course). SpeechDat was one of the first large scale projects that popularized the concept of 'phonetically rich speech material' that was already used in the design of the TIMIT corpus (Lamel et al., 1986), i.e., small sets of utterances designed to comprise all phonemes of a language in as many phonetic contexts as possible. The idea behind the phonetically rich utterances was that they should facilitate the training of truly context independent sub-word models.

In the eighties the ASR research community had come to the conclusion that there was no way around probabilistic models, and that the linguistic phonetic theory behind these models had better be as simple and global as possible. The success of Hidden Markov Models, trained with (by the standards of the time) large amounts of speech for which only an orthographic transcription was available, had convinced the research community that it would be much more cost-effective to collect very large amounts of coarsely annotated speech than to continue on the path explored by TIMIT, where fine phonetic transcriptions were provided. SpeechDat clearly reflects this trend: The design is completely oriented towards increasing the size of the corpora, while keeping the level of detail in the annotations to the minimum that was considered sufficient at that time. From the very beginning the SpeechDat partners have been aware that the quality of the annotation was one of the major factors determining the value of a corpus. Therefore, SpeechDat has designed and implemented detailed annotation guidelines and quality assessment procedures to monitor the quality of the annotations.

This paper is structured in the following way. Section 2 briefly presents the nature and chronology of the various SpeechDat projects that are mentioned throughout the paper. It should help in understanding the annotation decisions that were made in course of time of the individual projects. In Section 3 we will sketch the general structure of the databases, with a special eye on the properties of the annotation files. Then, we will deal with the conventions employed for the orthographic transcriptions of the speech (Section 4). Next, we will go into the properties of the annotation tools that were built for the projects (Section 5). The validation of the SLR within the SpeechDat projects is discussed in Section 6.

All SpeechDat reports referred to in this paper are public and can be accessed via http://www.speechdat.org/, by clicking the project and selecting the public deliverables. Information about contact addresses to obtain annotation tools can be found at the same address (see also Section 5). Most of the SpeechDat SLR are distributed through the European Language Resources Association ELRA (http://www.icp.inpg.fr/ELRA/cata/tabspeech.html). All SLR produced in the SpeechDat projects become publicly available, at the latest, 18 months after project end.

2. Overview of the SpeechDat Projects

Although not a direct descendant of the ESPRIT Project SAM, the SpeechDat concept is undoubtedly strongly influenced by the SAM project. The SAM project was conducted between 1989 and 1992 in the framework of ESPRIT II (as Project 2589). SAM is an acronym for "Multilingual Speech Input/Output Assessment, Methodology and Standardisation". The project aimed at defining standards, tools, and test protocols for speech data collections in all languages of Europe (SAM, 1992). The experiences gained in the SAM project and during the development of the

Dutch Polyphone corpus were of great value, when the European Commission decided to support the creation of multilingual telephone SLR. This resulted in the SpeechDat(M) project, which started in 1994 and was concluded in 1996. Within SpeechDat(M), 1000 speakers were recorded over the telephone for each of eight languages.

The follow-up project was SpeechDat(II), which covered 21 languages, including minority languages and language varieties. This project started in March 1996 and ended in December 1998. In SpeechDat(II) 28 SLR were created: 20 databases over the fixed network comprising 500 to 5,000 sessions from different speakers, 5 databases over the cellular network each comprising 1 session from 1,000 speakers or 4 sessions from 250 speakers, and 3 speaker verification databases comprising 1,000 sessions from 20 speakers, or 2,400 sessions from 120 speakers, respectively. The SpeechDat(II) project is described in Draxler (2000), Draxler et al. (1998), Höge et al. (1999), and Lindberg et al. (1998). Because of their common design the SpeechDat SLR are particularly suited to compare the performances of an ASR system for several languages (see Höge et al., 1999; Lindberg et al., 2000).

In the spirit of SpeechDat three new projects were launched. SpeechDat Car started in April 1998 and ended in January 2001. The project aimed at collecting 600 sessions per database recorded in the automobile, using four wide-band channels (60–7,000 Hz) on an in-car platform, and simultaneously recorded (on an independent platform) after transmission through the cellular GSM network. Basically, the same items were recorded as in the SpeechDat parent project, but the number of application words per session was increased substantially, in order to include a comprehensive set of words to control car equipment, and the handsfree use of the mobile telephone. Within SpeechDat Car, SLR were collected for nine EU languages: Danish, English, Finnish, Flemish/Dutch, French, German, Greek, Italian, and Spanish. For details the reader is referred to Sala et al. (1999), and Van den Heuvel et al. (1999). The extension to include parallel recordings, most of which are wide band, has had only a minor impact on the annotation procedures, tools and standards developed in the parent project. At about the same time SpeechDat East started out with telephone speech data collections for five central and eastern European languages over the fixed network, closely following the SpeechDat(II) specifications. The languages involved are Czech, Slovak, Polish, Hungarian (all 1,000 speakers), and Russian (2,500 speakers). The end date of this project was 30 November 2000. Finally, since the Summer of 1998, a similar enterprise is being carried out for four countries in Central and South America in the framework of the SALA project (Speechdat Across Latin America). SLR for 1,000 speakers will be created for the Portuguese variant spoken in Brazil, and the Spanish variants spoken in Mexico, Columbia, Venezuela, Argentina and Chile. Details about this project can be found in Moreno et al. (1998). The SpeechDat annotation conventions required only minor adaptations to accommodate the needs of the SpeechDat East and SALA languages.

A number of SLR was developed outside the aforementioned consortia. Nonetheless, they were built in close agreement to the design and annotation specifications of the SpeechDat projects. Examples include two SLR for Austrian German (Baum et al., 2000) and one for Australian English (http://www.callbase.com).

3. Structure and Format of the Databases

In this section we will globally describe the database formats used in the SpeechDat projects and the underlying principles to some extent (Section 3.1). Next, we deal in more detail with the contents of the annotation files (Section 3.2), which contain, among other things, the orthographic transcriptions of the speech, to which we shall turn in Section 4.

3.1. General Set Up

All SpeechDat databases have the same technical structure. The basic format of the ESPRIT Project SAM is followed (SAM, 1992). The influence of SAM is especially evident in the way in which speech signals and annotations are formatted and stored. Speech is stored in data files containing only the signal waveform samples without a header. All speech files as collected over telephone lines are in A-law (for SLR collected outside Europe also in Mu-law). An associated ASCII annotation (or label) file provides the transcription and other annotational information. In a sense the SAM structure is redundant, and thus wasteful. For instance, attributes pertaining to a full session are repeated in all annotation files for all speech recordings within that session. However, the size of the annotation files is negligible, compared to the storage capacity needed for the speech files.

The SpeechDat directory structure is independent of the content of the speech files. It does not contain any semantics regarding e.g. speaker or recording environment characteristics. Thus, it allows a fully automatic creation of a file system during recordings. Further, file names are designed such that they are unique even without the preceding directory pathname. Therefore the directory path of a file can be fully reconstructed from the file name and its extension. The directory tree itself has two functions. Firstly, it allows the storage of the recorded files of one call in a unique directory, which makes the databases much more transparent, and, secondly, the directory tree serves as a rapid search mechanism for the computer operating system software.

As mentioned above, each annotation file and corresponding speech file in a recording session are stored in the same directory, but in different files. The SpeechDat projects adopted the SAM standard in this respect and rejected the alternative of creating one file containing the speech preceded by an ASCII header partition with label information, such as in the NIST format. The latter format has significant disadvantages: 1. the need to create a special purpose tool to edit the label information in the header without interference with the speech part in the file (whereas in a split file approach, any simple text editor can be used to modify the annotation); 2. the obligation to store the speech on hard disk as well, if only the label information is needed (or to find or write special tools to split both types of data). On the other hand, it could be regarded a disadvantage of the SAM speech files that the basic signal coding information (sampling frequency, quantization resolution) is absent. This type of information is obligatory header information in the NIST SPHERE format, and has the benefit that speech files can always be decoded. For the SpeechDat projects this was not considered a serious problem, since all speech files in a database have the same signal coding specifications.

For the SpeechDat Car project, two different types of recordings were made. The four channel in-car recordings (with 16 kHz sampling frequency and 16 bit resolution for each channel) were stored in a multiplexed file, one for each corpus item. The second type was recorded through the cellular GSM network and stored in another file as A-law signal (with 8 kHz sampling frequency and 8 bit resolution). For both the multiplexed speech file and the A-law speech file a separate annotation file was created. This was necessary because they contain partly different information (like signal coding information and recording place), and it was time-efficient, because the annotation files could be made during recording. As a result, there are four files for each corpus item in a recording. These only differ in the final character of the file extension, which uniquely indicates the file type.

Each SpeechDat database comes with four types of accompanying files. (1) The lexicon file contains all the words found in the orthographic transcriptions together with a canonical phoneme transcription and, optionally, some alternative pronunciations. All phoneme transcriptions are in SAMPA (Wells, 1997). SAMPA is an acronym for SAM Phonetic Alphabet (SAM, 1992). We will return to this lexicon in more detail in Section 4.4. (2) Another file lists the following statistics about the signal acoustics for each speech file in the database: the maximum and minimum sample value, the mean sample value, the clipping rate and the signal to noise ratio (SNR). The tool that computes these statistics was developed by SPEX with the explicit aim to detect bad recordings. Extreme high clipping rates and low SNR values are considered good indicators for these types of recordings. (3) The contents list file, and (4) the speaker and recording table files contain selections of attribute values directly copied from the annotation files. As a consequence, these list and table files do not contain more information than can be retrieved from the label files (and are thus fully redundant). However, the list and table files contain all relevant information of a speech file in a relatively compressed form. This permits a rapid search for information without the need to access all annotation files. A full description of the database formats can be found in Senia (1997), and for SpeechDat Car in Draxler (1999).

3.2. Annotation Files

Annotation (or label) files adhere to a slightly modified SAM label format (SAM, 1992). Original SAM label files are full ASCII; they consist of lines, beginning with a three letter label mnemonic, a colon and a white space followed by field values separated by commas, up to a maximum of 80 characters per line.

SAM label fields can be either free-form text, single items from a fixed vocabulary, or lists of attribute-value pairs. Table 1 provides a list of mandatory common label mnemonics for all SpeechDat projects, i.e. labels that should be used in all label files for all SLR in the projects. For every label the type of information held

Table 1. Obligatory labels in the SpeechDat annotation files, their types, and their hierarchical scopes.

Label	Description	Type	Scope
LHD	Format name (SAM) + version	Format	*Full database*
SAM	Sampling frequency	Signal	”
SNB	Number of (8-bit) bytes per sample	”	”
SBF	Sample byte order (meaningless with single byte samples, “SNB: 1”)	”	”
SSB	Number of significant bits per sample	”	”
QNT	Type of quantization	”	”
DBN	Database name	Database	”
VOL	Database volume ID	”	*Multiple sessions*
SCD	Speaker code	Speaker	”
SEX	Speaker sex	”	”
AGE	Speaker age	”	”
ACC	Speaker accent	”	”
REG	Region of call	Environment	”
ENV	Calling environment	”	”
NET	Telephone network	”	”
PHM	Telephone hand set model	”	”
CCD	Corpus code of the item	”	”
SES	Session number	Session	*Single Session*
DIR	Signal file directory	”	”
REP	Recording place: city, country	Recording	”
RED	Recording date	”	”
RET	Recording time	”	”
SRC	Signal file name	”	*Speech file*
BEG	Labelled sequence start position	Time	”
END	Labelled sequence end position	”	
LBR	labelling during recording: begin sample, end sample, gain, min, max, orthographic text prompt	Utterance	”
LBO	Orthographic labelling: Begin sample, centre sample, end sample, orthographic transcription text	”	”

by it is printed as well. Four categories of labels with a different hierarchical scope, also shown in Table 1, are treated on a par in the label files, in the sense that they all appear in every annotation file of a database:

1. Labels with values which are typically identical throughout a complete SpeechDat database;
2. Labels which typically have identical values for a subset of the sessions in a database (multiple sessions), thus reflecting the design of the database (e.g. targeted distributions of speaker characteristics and recording environments);
3. Labels with values that typically apply to a single session only
4. Labels with values which are typically unique for one specific speech file in a database

For SpeechDat Car an extra set of labels was defined to represent the characteristics of in-car recordings. These attributes are listed in Table 2.

In SpeechDat Car, most of the SAM labels are automatically generated by the in-car platform software using the different fields that are filled by the operator at the beginning of each recording session. The

Table 2. Special obligatory labels used in the SpeechDat Car project, their types, and their hierarchical scopes.

Label	Description	Type	Scope
NCH	Number of channels	Signal	*Full database*
SPP	Speaker position	Environment	”
CAR	Car make and type	”	*Multiple sessions*
SCC	Scenario code	”	”
WTC	Weather condition	”	”
CEQ	Car equipment	”	”
MIP ”	Microphone position for each recording channel	Recording	”
MIT ”	Microphone type for each recording channel	”	”
EXN	Experimenter name	Session	”
LB {0 | 1 | 2 | 3} ”	Orthographic transcription for each in-car recording channel 0, 1, 2, 3	Utterance	*Speech file*
SYN ” ”	Synchronization mark; time between end of last DTMF ID code and end of speaker prompt beep	Signal	”

general principle is to allow as little freedom as possible in filling in the label fields to prevent editing errors, and to have meaningful label field entries that can be read by humans as well as machines. This means that mnemonic forms are used for items from a fixed vocabulary, e.g. STOP_MOTOR_RUNNING, HIGH_SPEED_GOOD_ROAD, etc. (as for label SCC), and also for attribute names and values, e.g. CLIMATE=ON, RADIO=ON, etc (as for label CEQ).

An example of an annotation file taken from the German SpeechDat Car database is printed in Fig. 1.

In addition, a large set of optional attributes is defined, but these are not described here, since they fit into the same annotational framework and thus do not add new information from an abstract formal point of view.

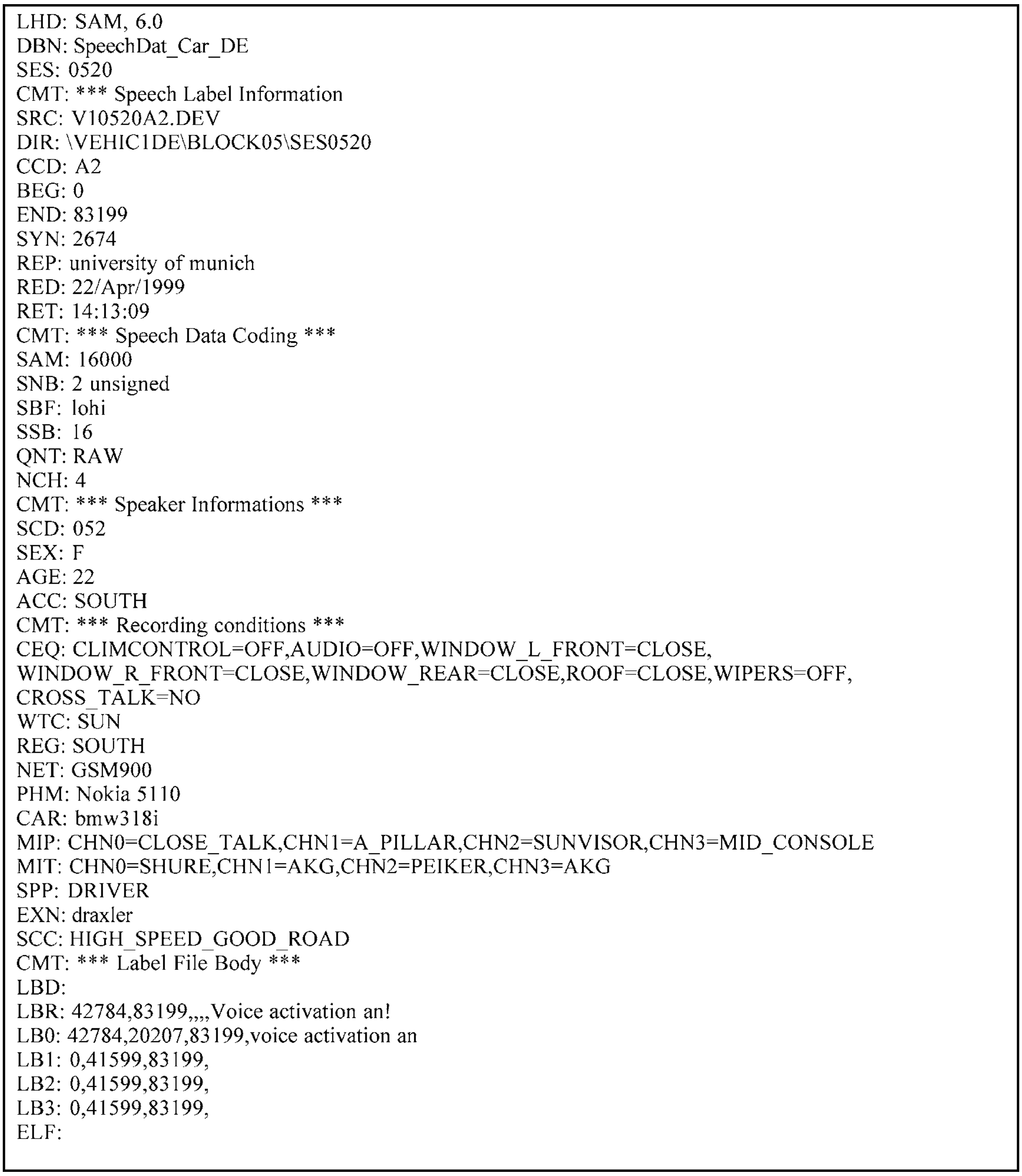
In retrospect, we regard the choice of the SAM format for data storage as a good decision. The SAM format has proven to be flexible enough to support all annotation needs. For example, it was quite simple to add new label mnemonics to meet the annotation needs for individual projects. Furthermore, since each and every label is repeated in each label file, the format permits an easy way to generate label files automatically, even on-line during recording (as in SpeechDat Car). The straightforward application of the three letter label mnemonic, followed by a value allows rapid generation of all kinds of meta-information files (tables and lists) by relatively simple software tools. The SAM format also allows an easy interface to the transcription tools (Section 5). Only the limitation to a maximum of 80 characters per line was felt as an obsolete restriction in the course of time and abandoned for this reason for the projects after SpeechDat(II).

We judge the SAM format as very efficient for similar SLR consisting of uncomplex, mutually unrelated speech utterances. As soon as more complex speech utterances are recorded (dialogues, longer monologues), higher level annotation layers are generally needed, together with efficient ways of connecting these layers, such as presented in Bird and Liberman (1999) and Mengel and Heid (1999).

4. Conventions for the Orthographic Transcriptions

The difference between a mere collection of speech and an actual speech database is “the fact that the latter is augmented with linguistic annotation (i.e. a symbolic representation of the speech),” as is attested in the EAGLES handbook (Gibbon et al., 1997:146). Without such a basic symbolic representation of the speech, the database becomes close to worthless.

The transcription of the speech utterances in the SpeechDat SLR is carried out solely at the orthographic level. First, we discuss the background principles underlying the transcription conventions (Section 4.1). Then, we present a brief overview of the transcription conventions and compare these to the EAGLES



```
LHD: SAM, 6.0
DBN: SpeechDat_Car_DE
SES: 0520
CMT: *** Speech Label Information
SRC: V10520A2.DEV
DIR: \VEHIC1DE\BLOCK05\SES0520
CCD: A2
BEG: 0
END: 83199
SYN: 2674
REP: university of munich
RED: 22/Apr/1999
RET: 14:13:09
CMT: *** Speech Data Coding ***
SAM: 16000
SNB: 2 unsigned
SBF: lohi
SSB: 16
QNT: RAW
NCH: 4
CMT: *** Speaker Informations ***
SCD: 052
SEX: F
AGE: 22
ACC: SOUTH
CMT: *** Recording conditions ***
CEQ: CLIMCONTROL=OFF,AUDIO=OFF,WINDOW_L_FRONT=CLOSE,
WINDOW_R_FRONT=CLOSE,WINDOW_REAR=CLOSE,ROOF=CLOSE,WIPERS=OFF,
CROSS_TALK=NO
WTC: SUN
REG: SOUTH
NET: GSM900
PHM: Nokia 5110
CAR: bmw318i
MIP: CHN0=CLOSE_TALK,CHN1=A_PILLAR,CHN2=SUNVISOR,CHN3=MID_CONSOLE
MIT: CHN0=SHURE,CHN1=AKG,CHN2=PEIKER,CHN3=AKG
SPP: DRIVER
EXN: draxler
SCC: HIGH_SPEED_GOOD_ROAD
CMT: *** Label File Body ***
LBD:
LBR: 42784,83199,,,,Voice activation an!
LB0: 42784,20207,83199,voice activation an
LB1: 0,41599,83199,
LB2: 0,41599,83199,
LB3: 0,41599,83199,
ELF:
```

Figure 1. Example of a SpeechDat Car annotation file (German).

recommendations (Section 4.2). Third, the alignment of the transcription symbols with the time course of the speech signal will be dealt with (Section 4.3). And, finally, we elaborate upon the implications of our approach with regard to the alignment of speech at the phoneme level by means of the lexicon (Section 4.4).

4.1. *Background Principles*

The level of the transcriptions in the SpeechDat projects is orthographic. The basic principles for the transcriptions are formulated in six points of departure (cf. Senia and Van Velden, 1997:5–6 and Gibbon et al., 1997:825–826) in what could be termed a manifesto:

1. The transcription is intended to be an ORTHOGRAPHIC, lexical transcription with a few details included that represent audible acoustic events (speech and non speech) present in the corresponding waveform files [...]
2. The transcription is intended to be a quick and broad transcription. Transcribers should not have to agonize over decisions, but rather realize that their transcription is intended to be a rough guide that others may examine further for details.
3. Transcriptions should be made in two passes: one pass in which WORDS are transcribed, and a second pass in which the additional details are added [...]
4. The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc.
5. The conventions comprize both mandatory and optional transcriptions. All transcriptions should precisely follow the mandatory guidelines. The optional transcriptions, if provided, should be documented and should follow these guidelines precisely. Markings which are optional have been chosen to be easily removed or translated by automatic means to yield the base transcription form.
6. The documentation provided with the database transcriptions should accurately provide details of which optional transcriptions were performed, and all relevant additional information, such as standard dictionary, preferred spelling variants, etc.

These principles are motivated by the vast amounts of data to be transcribed in the SpeechDat projects. The degree of detail and the correctness of the annotation of a database determines the value of the database, but both have a direct impact on the production costs. Each database from any of the SpeechDat family projects has an enormous quantity of data to be orthographically transcribed. A typical SpeechDat(II) database of 5,000 speakers contains 200,000 different speech files to be transcribed. In a SpeechDat Car database the total number of utterances pronounced by the speakers is larger than 70,000. In order to keep production costs and processing time within acceptable limits, it was decided that the orthographic representations should contain all the words spoken in the speech file, together with a very limited set of markers which enable the user to select roughly and rapidly subsets of speech files suitable for training or testing a speech recognizer. On the other hand, the transcriptions should be suitable to serve as a starting point to augment the transcriptions with other levels of annotation, e.g. phonemic information and segmentation.

4.2. Transcription Guidelines in SpeechDat

The transcription conventions agreed upon in SpeechDat were adopted from the conventions used by LDC/ARPA in producing the ATIS CD-ROMs (ATIS stands for Air Travel Information System). These were simplified considerably for the transcription of the Macrophone corpus, and later for the Dutch Polyphone corpus. In general, the SpeechDat transcription conventions come closest to those used in Polyphone, but some further simplifications were implemented to speed up the transcription process without loosing relevant information. Only the obligatory transcription conventions in the SpeechDat projects are listed below.

- The transcription is orthographic using in principle the words as they occur in the ‘reference’ dictionary of a language;
- Punctuation to denote sentences and clauses (viz. full stops, colons, semi colons and commas) are not used; only word-internal punctuation symbols, like apostrophes and hyphens, are allowed;
- Omissions from the prompt text in the actual speech utterance are not marked as such in the transcriptions. Missing words are just absent in the transcriptions;
- Self corrections (verbal deletions) are transcribed as normal text, e.g. “we meet in Florence I mean Venice”;
- Digits and numbers are written in full; long numbers may be split by blanks to avoid exponential growth of the lexicon;
- Special markers are attached to words that are mispronounced, or truncated during recording, or distorted during GSM transmission; no effort is devoted to indicate the degree of distortion, because this would take too much transcription time. E.g. a mispronunciation of *transportation* as *transportetation* is transcribed as “*transportation”
- Word fragments are considered mispronunciations (or recording truncations); also here transcribers should not try to indicate the exact cut-off point. Just the marker suffices. Everything more is considered a waste of transcription time.
- A special symbol is defined for unintelligible stretches of speech;
- A special symbol is defined for each of the following “background noise” categories: 1. filled pause

(e.g. 'uh', 'uhm'), 2. other speaker noise which is not speech (e.g. laughing, coughing), 3. stationary noise, 4. intermittent noise, and (in SpeechDat Car) 5. DTMF tone. These special symbols are placed at the location where the noise occurs (or starts, in case of stationary noise), but without splitting up words; in case a noise occurs or starts in a word, the symbol is put before the first word affected.

In comparison, the ATIS conventions (Shriberg et al., 1993) have a much larger symbol catalogue to capture background noises, speech overlaps, and the durations of noises. Further, they have more elaborated notations to transcribe mispronunciations, verbal deletions and word fragments. In contrast to SpeechDat, the Macrophone conventions (Taussig, 1997) include the use of a restricted set of abbreviations (mr, mrs, ms); the transcriptions of word fragments and stutters are cut at the place of the truncation; a larger set of background noise symbols is defined; and co-occurrence and time spans of background noises are annotated. The differences between SpeechDat and Dutch Polyphone are smaller, the most salient differences being the larger set of background noise symbols defined in Dutch Polyphone and the convention in Dutch Polyphone that mispronunciations and word fragments are not indicated in the transcription but reflected in the assessment of the recorded item as a whole.

A comparison with the transcription recommendations published in the EAGLES handbook (Gibbon et al., 1997:170–172) shows that the SpeechDat transcription conventions follow these recommendations to a large extent, but deviate in some points. The most important deviations in the SpeechDat transcriptions are:

- there are no explicit rules for the transcription of reduced word forms;
- only one symbol for filled pauses is defined;
- assessments of speech quality at item level or session level are not obligatory.

The transcription of a number of phenomena is optional in the SpeechDat SLR. But, if transcribed, there are some fixed conventions, stating how to deal with these phenomena. These conventions apply to e.g. the assessment of the speech quality at item level, and incidental prolongation of speech sounds. Optional transcription rules were also formulated for language dependent phenomena (e.g. *liaison* in French, see also Section 4.4).

4.3. The Link between the Orthographic Transcription and the Speech Signal

The connection between the transcribed orthographic string and the speech signal is brought about by several labels in the annotation file. The SRC label contains the name of the speech file which corresponds to the given transcription. The BEG and END labels contain the numbers of the first and last sample in the file. Thus, an exact segmentation is provided at utterance level, but at this level only.

In terms of the annotation graph model presented in Bird and Liberman (1999) the SpeechDat annotation framework is simple and straightforward. We can construct only one arc for the orthographic transcription offered in a SpeechDat label file. This arc has as label the utterance transcription and only two edge nodes (0 and 1) with the sample numbers denoting begin and endpoint of the file, respectively. Further information as to the alignment of speech and transcription is only provided by the left-to-right sequence of word order, but additional timing and segmentation information is not available in the annotation. The special symbols for noises can be regarded as fitting in the same left-to-right sequence as the words, since these are located in the string at the spot (before the word) where the noise occurs. Simultaneous events (like sounds starting in a word, or propagating through a word) cannot be distinguished from events between words in the SpeechDat transcriptions.

SpeechDat SLR do not provide segmentations at word or at phoneme level.

4.4. Correspondence between Lexicon, Orthographic Transcription and Speech

Apart from the alignment of the speech signal with the orthographic transcription, the alignment of the utterance at phoneme level is important for many ASR applications (Draxler, 2000). The CRIL (Computer Representation of Individual Languages) conventions present the proper framework to discuss these different types of annotations. The CRIL conventions were issued by the International Phonetic Association (IPA) in Kiel, 1989, and they introduce three systematically different levels for what could be called the text of a spoken utterance (Gibbon et al., 1997:152):

1. Orthographic level. This level contains the orthographic representation of the spoken text.

2. Phonetic level. This level specifies the phonetic form of a given word in its full (unreduced) segmental form, i.e. the citation (or canonical) form.
3. Narrow phonetic level. This level gives the narrow phonetic transcription of the words that were actually spoken. It is only on this level that phonetic categories can be directly related to the speech signal itself.

Transcriptions on the first and second level of the CRIL conventions are provided as standard for each SpeechDat corpus. The orthographic transcription is supplied in the annotation files and the canonical phonetic (or better, phonemic) transcription is provided by the lexicon. The lexicon that is delivered with each SpeechDat database can be considered as part of the annotation, if only because each word in the orthographic transcription can be replaced by its phonemic counterpart from the lexicon. The match between these is in principle perfect, since each word in the transcription should occur in the same spelling in the lexicon. Thus, the second level of the CRIL conventions can be established.

The aim of the SpeechDat transcriptions has never been to attain a narrow phonetic level as defined in the CRIL conventions. Obviously, the provision of an orthographic transcription together with a lexicon with canonical pronunciations does not even come close to such a target (Draxler, 2000:180–184). It is evident that a word in the actual utterance need not be pronounced as described in the lexicon. In running speech the canonical pronunciation is typically not realized, but individual phonemes are substituted, inserted, or deleted due to all kinds of reduction and coarticulation phenomena operating within and across words. These phenomena lead to alternative pronunciations of a word and their impact is not trivial, and, for that reason, much effort is spent to model such phenomena in ASR (Strik and Cucchiarini, 1999).

The point of departure in SpeechDat was the existence of electronic pronunciation dictionaries for the languages or language variants for which a corpus was collected. Through such a phonetic dictionary with canonical transcriptions an initial training of acoustic models is always possible. However, phonemic dictionaries are not available for all language variants. This can be due to the fact that proper SAMPA symbols have not yet been defined, and/or because such a lexicon is simply not available in any phonemic representation for the language variant.

It is interesting to explore how the orthographic transcriptions and the canonical lexicon supplied with each database can provide the basis for enhancing the annotation with alternative, more realistic, phoneme transcriptions. One way to cope with alternative pronunciations of a word is to allow such alternatives to appear in the lexicon. However, there is still no way of knowing which pronunciation variant was spoken in a specific utterance. This problem can only be solved by marking the correct pronunciation variant already *at the orthographic transcription level* (e.g. by attaching numbers to the words), so that the matching pronunciation can be retrieved from the lexicon. However, this may not solve all problems either. If the pronunciation variants are added by hand, the procedure becomes very expensive and unavoidably inconsistent. If the pronunciation alternatives are generated automatically by a set of standard phonological rules of a language, which are then aligned with the speech signal by an automatic forced recognition procedure, then overgeneration and especially undergeneration may still provide an inappropriate model of actual pronunciation. A data-driven approach (cf. e.g. Cremelie and Martens, 1998; Kessens et al., 2000) may remedy the aforementioned problems, but the resulting phoneme transcriptions may still not be of sufficient quality.

There are specific cases in which the realization of a pronunciation variant depends on the (immediate) context of a word. An example is liaison in French. Part of the solution, as implemented for French in SpeechDat, is to put an /h/ symbol for every phonemic transcription of words that orthographically start with an 'h', in order to indicate that the words resist liaison. Furthermore, higher level rules could be used during training and testing to predict where liaison will occur or not. The SpeechDat orthographic transcription conventions also have the option to put '+' after a consonant, if liaison is not applied, e.g. 'un petit+enfant' would indicate that the /t/ of *petit* is not pronounced. Finally, an additional annotation tier could be added in which the realized pronunciation variants are indicated by the annotators. In fact, such a solution was chosen for Czech in the SpeechDat East project.

5. Annotation Tools

Manual transcription is very valuable but is time consuming and very expensive. The SpeechDat family databases require a manual annotation of the whole database. Monitoring such a large quantity of data requires good annotation tools. Human errors must be avoided as much as possible and the designed tools must be fast to save time and costs, and robust to avoid errors. Ideally, the user just verifies the transcription proposed by the tool and validates it by a single keystroke.

There exists a rich variety of annotation tools generated in the various projects (Contantinescu et al., 1997; Bonafonte et al., 1998). The large number of annotation tools has come about because many researchers had constructed their own tools before they started collecting speech data within the SpeechDat framework. These tools have been modified and improved due to new specifications, or to the gained experience.

The designed tools for annotating SpeechDat databases have to fulfil the following requirements:

- Audio files can have different formats which the tool should be able to process. The recordings are typically made through fixed or cellular telephone networks (8 bit logarithmic A-law or Mu-law), or directly over microphone, as in SpeechDat Car (16 bit linear, sampled at 16 kHz).
- Typically the contents of the databases are composed of isolated numbers, connected numbers, names, words and sentences, and can be read or spontaneous. The transcription tools are designed to make a reliable transcription while taking advantage of the kind of contents they manage. They are provided with specific features, as shown in Section 5.1.4, to speed up the transcription procedure.
- The transcription tools can create and modify SpeechDat label files. In the SpeechDat framework a sheet identification number determines the contents that the speaker should have read. The task of the transcription tool is to create a new label file on the basis of this number (Fonollosa and Moreno, 1998). In the SpeechDat Car project, the contents of the prompts and the data from the speaker is known at the time of recording, so that the label files are already created during recording in the car. The transcription tool then only needs to access and modify the existing label file.
- All SpeechDat projects output slightly modified standard SAM label files as specified in each of the projects.

The annotation tools adopted in the SpeechDat projects have evolved during the different projects. Most of the tools are publicly available. A number of them are freeware like some versions of WWWTranscribe (Draxler, 1998); others can be licensed like Vox!, Annotator (Contantinescu et al., 1997), NaniBD (Nogueiras and Moreno, 1998), and JavaSgram (Cristoforetti et al., 2000).

5.1. Features of Existing Tools

5.1.1. Platforms. Most of the annotation tools work on Windows 95/98/NT, UNIX (Linux, SunOS, SCO, IRIX) or are platform independent. An often used software environment is Perl and JAVA, but GNU/ncurses, MATLAB and C are also used.

5.1.2. Transcription Procedure. The tools are user friendly, since they have to be used by non-experienced transcribers. The procedure must be as simple as possible. All tools automatically choose the signals and speakers to annotate, play the audio signal and show the prompted text in an editing window. The prompted text should need as few modifications as possible to minimize errors. The tools typically use single key strokes or special button panels to insert the noise markers at the proper location in the transcription string. Some tools distinguish the noise markers from the text by colours or highlights. The tools automatically change them to the text strings defined in the database specifications. The transcriber may hear the signal as many times as he or she needs, modify the text, add the necessary noise marks and validate the corrections.

5.1.3. Graphical Display. The screen display must be easy to understand. Most of the displays have an area to display the waveform, an area with push buttons to make the annotation easier and an area to edit the transcription. It has been observed in the last releases of the annotation tools that it is easier and faster for a minimally trained transcriber to use keyboard shortcuts instead of the mouse or click buttons to carry out the annotation task. Moreover some actions can be joined in just one key stroke: Save/Select/Display/Play. The latest versions of NaniBD, WWWTranscribe and Vox! all prefer the use of the keyboard instead of the mouse for most parts of the annotation task.

5.1.4. Additional Functions. A common feature of the annotation tools is their ability to run some specific programs:

- Convert lower to upper case
- Convert upper to lower case

- Convert digit strings, numbers or strings of numbers to their orthographic representation
- Convert dates in numerical form to the corresponding orthographic form
- Convert spelling strings to the corresponding strings of letter names

These actions can be done off-line or on-line. The preprocessing of the prompt text into a proposed transcription text, prior to the actual transcription work by the annotator, is a typical off-line procedure. Similarly, all kinds of modifications of the transcribed text afterwards (like case conversion, symbol conversions to the correct format) are off-line procedures. On-line procedures are developed to facilitate and speed up and check the transcription of all the items containing numbers, dates and spellings (see Section 5.1.5 below). A single keystroke can be used to perform all the checks simultaneously.

Another aid to the transcribers is the option of having rapid access to the information in the label files describing all the characteristics of the edited speech file (e.g. noisy condition, recording date, speaker age, etc).

5.1.5. Dictionaries. Various tools (Vox!, NaniBD, WWWTranscribe) incorporate dictionaries. We can distinguish two types: the first type is a dictionary designed for short, typically single word utterances. The transcriber can choose one word from a finite set. The tool reduces spelling errors while annotating words from a short finite set (cities, forenames, surnames, company names, keywords...). These dictionaries can also be used for other purposes. For instance, a common question in the recording of databases is the town or region where the subject grew up. The answer is used to determine the accent of the speaker. The dictionary helps to find the correct spelling of the name of the town or region, the program checks simultaneously the dialectal region and writes the result in the corresponding file (Nogueiras and Moreno, 1998).

The second type is a spelling dictionary that is applied to all sentences. Mismatches can either be shown to the transcriber or stored in a separate file and processed later (Draxler, 1998).

5.1.6. Integration of ASR Systems. The annotation tools have evolved in the sense that most of them can execute external commands and are designed to easily include new external commands. ASR software can be integrated to work on-line in some tools, e.g. Annotator and NaniBD. The annotation conventions of the SpeechDat family databases do not require segmentation markers (or boundaries) between phones, words or any other units. A recognition system to initialize these boundaries is therefore not necessary, but experience shows that it is convenient to run ASR software off-line to prepare a first transcription text for the human annotator, so as to minimize actual transcription time. This automatic preparation of the transcription text by ASR technology is of specific value for:

- Key numbers: if what the speaker reads is keyed by a number, a previous identification of this number is necessary to trigger the original text. Many tools automatically recognize the number and prepare label files with these expected texts automatically before the actual transcription session starts.
- Spontaneous answers such as yes/no questions, forenames, surnames and city names.
- Spontaneous telephone numbers
- Strings of numbers spontaneously grouped. Read strings of numbers such as credit card numbers, telephone numbers or PIN codes, allow for various groupings of numbers. For instance, the digit sequence ‘1 3 3 4’ may be pronounced as ‘one three three four’ or ‘one thousand three hundred thirty four’ or ‘thirteen thirty four’, etc. ASR techniques are applied to detect what groupings have been produced and to show the transcriber the numbers that the speaker really uttered to facilitate the transcription task.

5.2. Annotation of Multichannel Corpora

The tools discussed so far are specifically developed for one-channel annotation. However, in SpeechDat Car five replicas of each utterance are recorded: a close talk microphone, three in-car hands-free microphones, and a microphone connected to a GSM-based recording system. As a consequence, in this case it is convenient to use an annotation tool which allows visualization and transcription of all the given channels at the same time.

5.2.1. Need of Independent Channel Annotation. When dealing with multichannel corpora, an important question concerns the alternative of either making annotations distinctly for each channel or applying annotation of one channel to the other ones. Multichannel SLR are typically recorded by using microphone arrays. These generally consist of a number of the same microphones, placed close to each other; alternatively, the microphones may be of different brands and placed at larger distances from each other. When two microphones are located far from each other and have different response characteristics, relevant discrepancies may be found in the corresponding recordings of the same acoustic event. Depending on the study or the application for which the corpus is collected, the resulting discrepancies may justify the need of an independent channel annotation.

In the case of SpeechDatCar (where only the annotation of the close-talk microphone is mandatory), the characteristics and positions of microphones may lead to very different recordings of the same speech utterance. In the following, a few examples are given to illustrate these aspects:

- A burst event, caused by rough road conditions, is more manifest in the signal recorded by the far-microphone close to the A-pillar than in the other microphone signals; and it may be not audible at all in the close-talk recording;
- A noise event produced by the speaker (e.g. lip smack) is probably only audible in the close-talk recording;
- A distortion event, due to GSM transmission problems, is obviously not present in the in-car recordings.

Clearly, it is desirable to have a tool which makes independent channel annotation easy and allows the transcriber to copy rapidly the same transcription to all of the channels and then modify the transcription for individual channels.

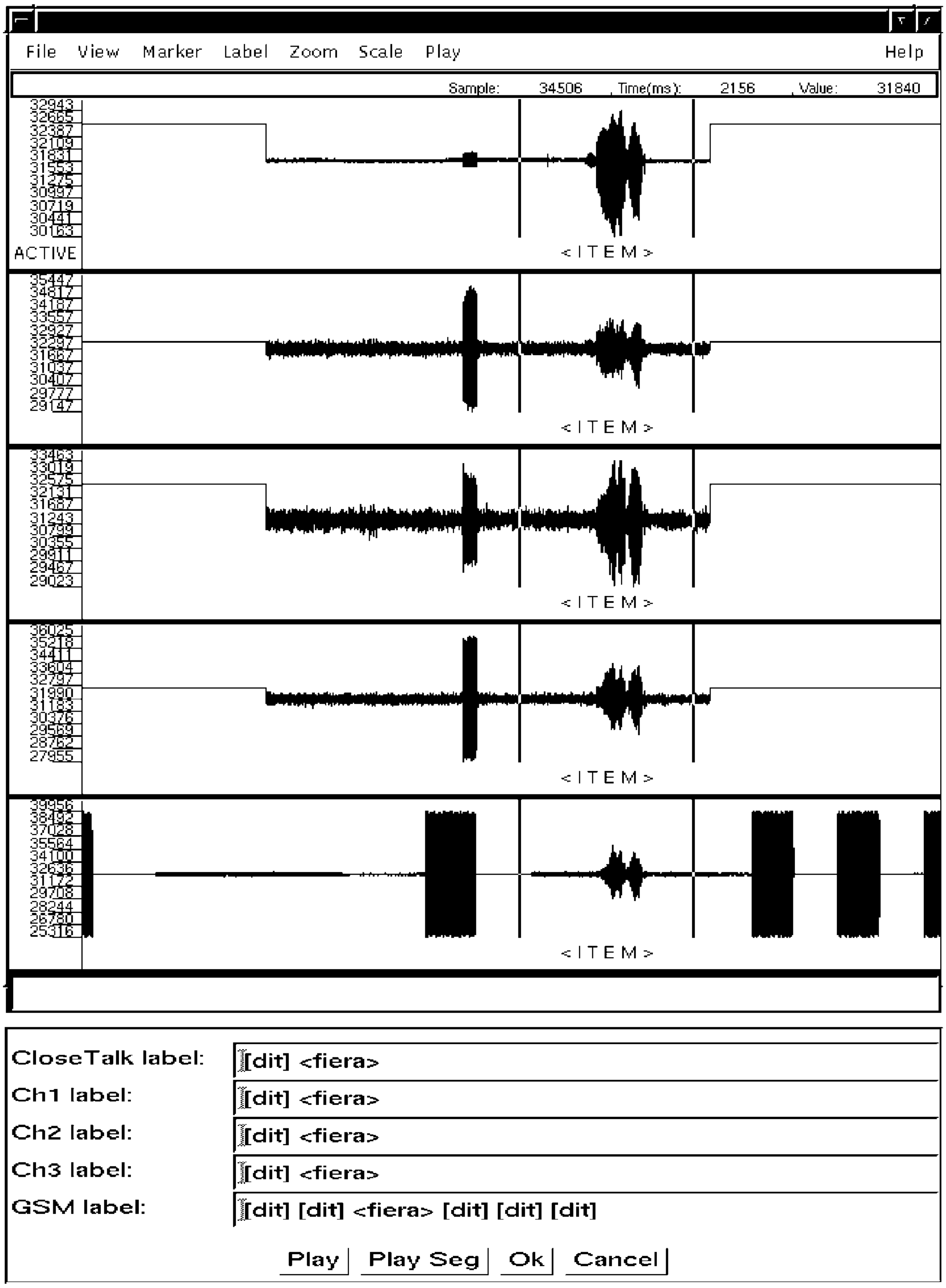
5.2.2. An Example of a Multichannel Annotation Tool: JavaSgram. JavaSgram is an example of a tool which has these extended features. JavaSgram was developed at ITC-IRST Trento, Italy, with the objective of being flexible for independent annotation of all of the input channels. The signals corresponding to these channels can be visualized, zoomed, and unzoomed together in a compact graphic representation (see Fig. 2). In this way, one can annotate only the reference channel, while visually checking the other channels to detect possible noise events (often corresponding to anomalous changes in the energy/signal envelope). Different multiplexed input file formats can be used (e.g. the raw format used in SpeechDatCar and the NIST-SPHERE format) with any number of input channels.

The tool provides fast access to and playing of different channels. It provides the transcriber with all the information needed for rapid successive annotation of all the items belonging to a given recording session. To further speed up this task, a mask can be defined that allows to initialize the annotation of every item with some constant labels, corresponding to events which are always (or very often) present in the given channels (e.g. DTMF tones). Furthermore, it is worth noting that markers can be introduced either for a specific channel or for all the channels to indicate beginning and end of acoustic events; either single or multiple independent labelling can be accomplished for a given speech segment. The tool was written in Java in order to ensure portability to different platforms and operating tools. To speed-up the annotation task, JavaSgram was enriched with a number of automatic pre/post-processing modules. Further, once the annotation of a given session has been completed, the tool can be re-started with a new configuration, specific for the verification phase; in this case all paths are re-defined and different functions are enabled.

5.2.2.1. Synchronization. All in-car and GSM channels of a recording are time aligned and displayed (see Fig. 2). For this purpose, the GSM signal is upsampled to 16 kHz and then converted to linear dynamics. Subsequently, a Crosspower Spectrum Phase based technique (Omologo and Svaizer, 1997) is applied to in-car and GSM data in order to align them automatically at a sample accuracy level. In other words, a spectral analysis is conducted on the close-talk and the 16 kHz-GSM signals as if they were two delayed versions of the same signal. Performances of this alignment module are excellent; in fact, for more than 99% of the items there is no need for manual re-alignment. Note that generally the GSM signal is longer than in-car signals; hence, while creating the multiplexed file a number of zeros is introduced at the beginning and at the end of the in-car signals.

Apart from a better visualization of all of the signals, another advantage of this synchronization procedure is that during annotation one can immediately detect typical GSM distortions, such as those due to loss of samples, which sometimes occur in the case of low quality GSM communication.

5.2.2.2. Segmentation. Segmentation of the speech portion in an utterance is not mandatory for the SpeechDat projects. However, to annotate where the



Sample: 34506 , Time(ms): 2156 , Value: 31840

CloseTalk label: [dit] <fiera>

Ch1 label: [dit] <fiera>

Ch2 label: [dit] <fiera>

Ch3 label: [dit] <fiera>

GSM label: [dit] [dit] <fiera> [dit] [dit] [dit]

Figure 2. Example of a screen shot of JavaSgram, showing five synchronized channels from a SpeechDat Car item. The upper window shows the signal from the close talk microphone; the four windows below it show the corresponding signals from three far-talk microphones and the corresponding GSM signal, respectively. The GSM channels are surrounded by DTMF tones, and the other channels have a prompt beep before the speech utterance. The bottom panel shows the corresponding orthographic transcriptions. The triangular brackets around *fiera* indicate that the text needs verification by the annotator; the brackets are not part of the final transcription.

speech portion of a recorded item starts and ends is very useful for training speech recognizers, especially in the case of noisy speech. Such a segmentation is shown in Fig. 2.

This type of segmentation requires substantial efforts for human annotators. On the other hand, the exact locations of the item boundaries are not critical. In the case of SpeechDat Car data annotation, a traditional segmentation procedure would not provide a reliable set of item boundaries, since it would include many events (such as DTMF tones) as part of the item. Related to JavaSgram a specific segmentation tool was developed, based on the use of a Spectral Variation Function (SVF) (Brugnara et al., 1993). In fact, SVF exhibits better properties than other energy-based features providing an effective representation of sudden spectral changes. This segmentation module was developed to exclude prompt beeps and DTMF sequences from the item portion, as well as possible short speaker "noise" or other distortion components sometimes occurring before the item (see Fig. 2). As a result of the application of this pre-processing tool, during manual annotation the proposed item boundaries are changed in less than 3% of the cases.

6. Validation

In the context of the SpeechDat projects the term "validation" is used to refer to the process in which a database is checked against the specifications and corresponding tolerance intervals (together "the validation criteria") that were agreed upon by the consortium members. The SpeechDat projects follow a unique validation campaign in order to assure that all databases meet the specifications. The campaign is unique in the sense that an independent organization which is integrated into the project itself checks all databases. This approach warrants that each database that is produced by the consortium is in agreement with a well-defined set of minimum quality standards (see below). An important motivation for this quality check is the free exchange policy of databases within each consortium. This policy requires guarantees that the databases are of equally good quality. The validation centre in all five SpeechDat projects is the Speech Processing EXpertise centre (SPEX). The validation criteria and procedures are described and evaluated in Höge et al. (1999) and Van den Heuvel (1997, 1999, 2000a). A more general overview of SLR validation is presented in Van den Heuvel (2000b).

A lot of the validation work is done automatically. Software was written to check file formats, internal consistency, missing files, transcription symbols used, speaker and environment balances, etc. The software also checks if all obligatory labels are used in all label files and if the label attribute values are correct and consistently used. The interpretation of the software output involves human intervention, as does the editing of the validation report. The automatic part of the validation is done on a Unix platform and the software is written entirely in Perl. The software consists of a part with subroutines that is the same for all equivalent databases and a database-dependent file in which values for variables are defined. Accordingly, only this latter file needs to be updated for the validation of a new database. The file also contains various variables referring to typical errors encountered in the past and can be used to avoid that the database validation terminates prematurely because of sometimes trivial errors in the database.

The validation of the transcription quality is carried out by a native speaker of the language of the database. He or she checks about 2,000 randomly selected utterances from the corpus. This number renders the transcription check sufficiently reliable from a statistical point of view. The validator does not make a new transcription from scratch, but rather departs from the given transcription, because the golden rule for the transcription validation is that the given transcription gets the benefit of the doubt. In this way, only overt errors are reported. This avoids debates with the database producer/owner about subjective details afterwards. The following criteria apply: A maximum of 20% of the checked items may contain an error in the transcription of the non-speech markers, and a maximum of 5% of the items may contain an error in the transcription of the speech.

In the SpeechDat(II) project about one out of six databases needed a revalidation. In the majority of cases, the reason for such a revalidation was one or more incomplete corpus items. Only in one case was the orthographic transcription of insufficient quality.

7. Conclusion

In this contribution we have presented the SpeechDat SLR in general, and their annotation conventions in particular. Special attention was given to the conventions for the orthographic transcriptions, the link between these orthographic transcriptions and corresponding phoneme transcriptions via the lexicon, the tools created to carry out the annotations, and the validation procedures implemented to safeguard the quality of the annotations. It was pointed out that the SpeechDat SLR deliberately have a very flat structure in almost every respect. The speech files and annotation (i.e. label) files are stored in redundant directories which are fully recoverable from the names of the files. The additional index and table files contain only (subsets of) labels which are also present in the annotation files. The labels in the annotation files are of different types and scopes, but are projected into the same flat level in the sense that they appear in each annotation file. These format specifications were adopted from the SAM standards, and have proven very valuable for our work. This demonstrates that the SAM standards could well be tailored to the annotation of a specific type of SLR that is considered elementary for the training and testing of various types of ASR systems. We consider this an innovative and encouraging application of the SAM standards in its own right.

The transcriptions of the speech are orthographic, and enriched by a limited set of additional symbols (for e.g. background noises and mispronunciations). Effort was put into reducing the transcription conventions to the bare minimum needed to successfully train and test ASR systems. These limitations keep the (manual) transcription task in manageable proportions.

The orthographic transcriptions can be transformed into phonemic transcriptions by means of the lexicon with SAMPA phoneme transcriptions that comes with each SpeechDat database. An obvious consequence of this approach is that a phonemic transcription of a word is the same for each occurrence of the word in the database, which needs not at all reflect its actual pronunciation in a given instance. Yet, accurate phonemic transcriptions are not needed for initial training and testing of an ASR system.

The SpeechDat projects produced a set of valuable annotation tools. They aim at assisting the human transcriber to realize a rapid transcription of the speech utterances by featuring quick search mechanisms and time-efficient shortcuts (e.g. push buttons for digits and background noise symbols). The SpeechDat Car SLR having multi-channel input pose specific problems which require additional functionalities of the annotation tools, e.g. the synchronization of the signals.

The quality of the annotation is checked within the SpeechDat projects by an independent validation centre. To this end the specifications and their tolerance margins need to be clearly formulated. This serves again as an aid to make the SpeechDat SLR as uniform and mutually compatible as possible.

Acknowledgments

SpeechDat(M) was supported by the EC under contract LRE-63314. SpeechDat(II) was supported by the EC under contract LE2-4001. SpeechDat-Car was supported by the EC under contract LE4-8334. SpeechDat-East was supported by the EC under contract INCO-Copernicus 977017.

References

Baum, M., Erbach, G., and Kubin, G. (2000). SpeechDat-AT: A telephone speech database for Austrian German. In *Proc. LREC'2000 Satellite Workshop XLDB—Very large Telephone Speech Databases*, 29 May 2000, Athens, Greece, pp. 51–56.

Bernstein, J., Taussig, K., and Godfrey, J. (1994). Macrophone: An American English telephone speech corpus for the Polyphone project. *Proc. ICASSP-94*, Adelaide, pp. 81–83.

Bird, S. and Liberman, M. (1999). A formal framework for linguistic annotation (Technical Report MS-CIS-99-01). Department of Computer and Information Science, University of Pennsylvania.

Bonafonte, A., Moreno, A., Draxler, C., Van den Heuvel, H., and Yli-Hietanen, J. (1998). Annotation tools (SpeechDat Car Technical Report SD3.1.2).

Brugnara, F., Falavigna, D., and Omologo, M. (1993). Automatic segmentation and labeling of speech based on Hidden Markov models. *Speech Communication*, *12*:357–370.

Contantinescu, A., Caloz, G., Draxler, C., Van den Heuvel, H., Sanders, E., Winsky, R., Nataf, A., Chatzi, I., Senia, F., Moreno, A., and Johansen, F. (1997). Report on developed tools (SpeechDat Technical Report SD3.1.2).

Cremelie, N. and Martens, J.P. (1998). In search of pronunciation rules. In *Proc. of the ESCA Workshop "Modeling Pronunciation Variation for Automatic Speech Recognition,"* Rolduc, pp. 23–27.

Cristoforetti, L., Matassoni, M., Omologo, M., Svaizer, P., and Zovato, E. (2000). Annotation of a multichannel noisy speech corpus. In *Proc. of the Second International Conference on Language Resources and Evaluation*, Athens, pp. 1547–1550.

Den Os, E.A. den, Boogaart, T.I., Boves, L., and Klabbers, E. (1995). The Dutch Polyphone corpus. In *Proc. Eurospeech-95*, Madrid, Spain, pp. 825–828.

Draxler, C. (1998). WWWSigTranscribe. A JAVA extension of the WWWTranscribe toolbox. In *Proc. of the First International Conference on Language Resources and Evaluation*. Granada, Spain, pp. 1313–1316.

Draxler, C. (1999). Specification of database interchange format (SpeechDat-Car Technical Report D1.3.3).

Draxler, C. (2000). Speech databases. In F. Van Eynde and D. Gibbon (Eds.), *Lexicon development for Speech and Language Processing*. Dordrecht, Boston, London: Kluwer Academic Publishers, pp. 169–204.

Draxler, C., Van den Heuvel, H., and Tropf, H.S. (1998). SpeechDat experiences in creating large multilingual speech databases for teleservices. In *Proc. of the First International Conference on Language Resources and Evaluation*, Granada, pp. 361–366.

Fonollosa, J.A.R. and Moreno, A. (1998). Automatic database acquisition software for ISDN PC cards and analogue boards. In *Proc. of the First International Conference on Language Resources and Evaluation*, Granada, pp. 1325–1328.

Gibbon, D., Moore, R., and Winski, R. (Eds.) (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin, New York: Mouton, de Gruyter.

Höge, H., Draxler, C., Heuvel, H. van den, Johansen, F.T., Sanders, E., and Tropf, H.S. (1999). Speechdat multilingual speech databases for teleservices: Across the finish line. In *Proc. EUROSPEECH'99*, Budapest, Hungary, 5–9 Sept. 1999, pp. 2699–2702.

Kessens, J.M., Strik, H., and Cucchiarini, C. (2000). A bottom-up method for obtaining information about pronunciation variation. In *Proc. of ICSLP 2000*, Beijing, China, pp. 274–277.

Lamel, L., Kassel, R.H., and Seneff, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. *Proc. DARPA Speech Recognition Workshop*, pp. 100–109.

Lindberg, B., Comeyne, R., Draxler, C., and Senia, F. (1998). Speaker recruitment methods and speaker coverage. Experiences from a large multilingual speech database collection. In *Proc. ICSLP 98*, Sydney, pp. 2731–2734.

Mengel, A. and Heid, U. (1999). Enhancing reusability of speech corpora by hyperlinked query output. In *Proc. Eurospeech 99*, Budapest, pp. 2703–2706.

Moreno, A., Höge, H., Koehler, J., and Marino, J. (1998). SpeechDat across Latin America. Project SALA. In *Proc. of the First International Conference on Language Resources and Evaluation*, Granada, pp. 367–370.

Nogueiras, A. and Moreno, A. (1998). NaniBD: A set of tools for transcribing and validating speech databases. In *Proc. of the First International Conference on Language Resources and Evaluation*, Granada, pp. 1359–1365.

Omologo, M. and Svaizer, P. (1997). Use of the cross-power spectrum phase in acoustic event location. *IEEE Trans. on SAP*, *5*(3):288–292.

Sala, M., Sanchez, F., Wengelnik, H., Van den Heuvel, H., Moreno, A., Le Chevalier, E., Deregibus, E., and Richard, G. (1999). Speechdat-Car: Speech databases for voice driven teleservices and control of in-car applications. In *Proc. EAEC 99*, Barcelona, pp. 90–98.

SAM (1992). User guide to ETR tools. SAM: Multi-lingual speech Input/Output Assessment, Methodology and Standardisation. Ref: SAM-UCL-G007.

Senia, F. (1997). Specification of speech database interchange format (SpeechDat Technical Report SD1.3.1).

Senia, F. and Van Velden, J. (1997). Specification of orthographic transcription and lexicon conventions (SpeechDat Technical Report SD1.3.3).

Shriberg, L., Price, P., Garofolo, J., and Fisher, W. (1993). ATIS. SR output (".sro") transcription conventions. http://www.ldc.upenn.edu/Catalog/readme_files/atis3/sro_spec.html.

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, *29*:225–246.

Taussig, K. (1997). Macrophone transcription. http://www.ldc.upenn.edu/Catalog/readme_files/macrophone/transcrp.html.

Van den Heuvel, H. (1997). Validation criteria (SpeechDat Technical Report SD1.3.3).

Van den Heuvel, H. (1999). Validation criteria (SpeechDat Car Technical Report D1.3.1).

Van den Heuvel, H. (2000a). SLR validation: Evaluation of the SpeechDat approach. In *Proc. LREC'2000 Satellite workshop XLDB—Very large Telephone Speech Databases*, 29 May 2000, Athens, Greece, pp. 40–45.

Van den Heuvel, H. (2000b). The art of validation. *ELRA Newsletter*, *5*(4):4–6.

Van den Heuvel, H., Bonafonte, A., Boudy, J., Dufour, S., Lockwood, Ph., Moreno, A., and Richard, G. (1999). SpeechDat-Car: Towards a collection of speech databases for automotive environments. In *Proc. of the Workshop for Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, pp. 135–138.

Van den Heuvel, H., Boudy, J., Comeyne, R., Euler, S., Moreno, A., and Richard, G. (1999). The SpeechDat-Car multiligual speech databases for in-car applications: Some first validation results. In *Proc. Eurospeech 99*, Budapest, pp. 2279–2282.

Wells, J. (1997). *Standards, Assessment, and methods: Phonetic Alphabets*. London: University College.

Web References

Speechdat Family: http://www.speechdat.org/
SpeechDat: http://www.speechdat.org/SpeechDat/
SpeechDat Car: http://www.speechdat.org/SP-CAR
SpeechDat East: http://www.fee.vutbr.cz/SPEECHDAT-E/
SALA: http://gps-tsc.upc.es/veu/sala/
ELRA: http://www.icp.inpg.fr/ELRA/home.htm