

**1. DELIVERABLE IDENTIFICATION**

Identification number	LE2-4001-SD1.1.3
Type	Technical Report
Title	Definition of corpus, scripts and standards for Speaker Verification
Status	Final
Deliverable	SD1.1.3
Work Package	WP 1
Task	Task 1.1
Period covered	T0 - T6
Date	10 February, 1997
Version	3.3
Number of pages	14
Author(s)	Kamran Kordi, GEC-Marconi Hirst, Aurelie Nataf, Matra, Andre Constantinescu IDIAP, Richard Winski, Vocalis
Workpackage (WP)/ Task (T) responsible	WP 1 - Richard Winski Task 1.1 - Kamran Kordi, GEC-Marconi Hirst
Project contact point:	Harald Hoege Siemens AG, ZFE T SN 5, D-81730 München phone: + 49 89 636 3374 fax: + 49 89 636 49802 e-mail: hoege@habicht.zfe.siemens.de
CEC project officer	Mr. J. Soler
Status	Public
Actual distribution	
Supplementary notes	Consortium and CEC

Key words	Telephone, speech, database, contents, design, speaker verification
Abstract	<p>This document provides a specification of the contents of the Speaker verification Database (SDB) to be collected over the fixed and mobile telephone network in 3 European languages. It should be read in conjunction with the accompanying deliverables in WP1 specifying recording conditions, speaker attribute coverage, transcription and validation.</p> <p>As the specification has been developed in consultations with all partners, and is intended to provide a common resource for developing general speaker verification services, the specification is identical or very similar in most respects for each language database. This is intended to ensure that the databases are very similar in terms of their exchange value and potential for speaker verification technology developments in each of the represented languages.</p>
Status of the abstract	public

Received on	
Recipient's catalogue number	

## 2. DOCUMENT EVOLUTION

Version	Date	Status	Notes
1.0	12/7/96	first draft	Discussed by WP1 partners, Athens
2.0	16/10/96	second draft	Discussed by SDB partners in Aalborg
3.0	9/12/96	pre-final	Discussed by e-mail
3.1	10/2/97	final for review	Discussed by e-mail
3.2	24/2/97	final	Final approved version

# Contents

<b>1. DELIVERABLE IDENTIFICATION.....</b>	<b>1</b>
<b>2. DOCUMENT EVOLUTION.....</b>	<b>2</b>
<b>3. INTRODUCTION.....</b>	<b>4</b>
<b>4. DATABASE CONTENTS.....</b>	<b>4</b>
4.1 OBJECTIVES.....	4
4.2 LANGUAGE COVERAGE.....	4
4.3 SWISS FRENCH POLYVAR DATABASE.....	5
<b>5. DESCRIPTION OF CONTENTS.....</b>	<b>5</b>
5.1 DIGIT/NUMBER STRINGS.....	5
5.1.1 <i>Credit card numbers</i> .....	5
5.1.2 <i>Personal Identity Numbers (PINs)</i> .....	6
5.2 SEQUENCE OF DIGITS.....	6
5.3 PHONETICALLY RICH SENTENCES.....	6
5.3.1 <i>Common sentences</i> .....	6
5.3.2 <i>Speaker-specific sentences</i> .....	6
5.4 SPELLED NAME / WORD.....	6
5.5 SPEAKER-SPECIFIC APPLICATION WORDS.....	7
5.6 SPEAKER-SPECIFIC NAMES.....	7
5.7 UTTERED AND SPELLED SPEAKER NAME.....	7
<b>6. RECORDING SPEAKER-SPECIFIC DETAILS.....</b>	<b>7</b>
<b>7. RECORDING ENVIRONMENT.....</b>	<b>8</b>
<b>8. SPEAKER RECRUITMENT.....</b>	<b>8</b>
<b>9. UTTERANCE ELICITATION AND PROMPT SHEETS.....</b>	<b>8</b>
<b>10. APPENDIX A.....</b>	<b>10</b>
<b>11. APPENDIX B.....</b>	<b>12</b>
<b>12. APPENDIX C.....</b>	<b>14</b>

### **3. Introduction**

This document details specifications of the contents of the Speaker Verification Database to be collected over terrestrial and mobile telephone channels by two partners in SpeechDatII project. The databases are intended to provide a common resource for 2 European languages, namely, British English and French, and a comparable database in Swiss French will also be made available in the SpeechDat project. However, this document can also serve as a guideline should other participants decide to collect their own databases at a later date.

The specification has been developed in consultation with all partners, and is intended to be identical in all major aspects for each language database. This ensures that the databases are very similar for the purpose of establishing a common exchange value and to provide the same potential for speaker verification technology developments in each of the languages.

This document should be read in conjunction with the accompanying deliverable reports SD1.2.3 (Environmental and speaker specific coverage of SDB), and SD1.3.3 (Annotation standards and validation criteria for SDB), and the related reports addressing the mobile telephone recordings and Fixed Network Database specification report SD1.1.1.

### **4. Database contents**

#### **4.1 Objectives**

The databases are intended to provide sufficient data for training of different classes of speaker verifiers, including text dependent and text independent systems. Test material is provided by seeding many of the relevant training items in the FDB and MDB contents specifications.

The broad technical properties of the databases are:

- 120 speakers: 60 male and 60 female.
- Speech samples collected digitally, directly from the telephone network in a-law.
- Each speaker makes 20 calls, with a minimum of 3 day interval between the calls.
- Broad regional accents of each language are covered in the collection exercise.

The vocabulary for the database contains:

- Isolated digits and continuous digit/number strings
- Phonetically rich sentences
- Application words, to simulate user-selected passwords
- Spelled names and words

The vocabulary used in the database collection will provide a realistic basis for training and testing speaker verification systems both in text dependent and text independent modes. To ensure a minimum of 120 speakers are recorded for the database, it is proposed that all partners should recruit 150 speakers initially in order to cope with the possible withdrawing of a number of speakers from the collection exercise.

#### **4.2 Language coverage**

The full database specification has as its goal to provide recordings of 120 speakers for two European languages, namely, French and British English. Each database will cover at least 2 broad accents for the language it is intended for. In each case vocabulary items are to be present in sufficient numbers to enable adequate training for different classes of speaker verification systems.

### 4.3 Swiss French PolyVar database

The Swiss PolyVar speaker verification database has been designed prior to the SpeechDat SDB databases, and does not follow the same specification as for English and French although it has similar content and design features. In particular it is based on recordings of 20 speakers making 50 calls each, and does not have the same correspondence of items with the FDB databases.

## 5. Description of contents

The following table provides a summary of utterances and their frequency per call for the speaker verification database. In each case the items are examples of read rather than spontaneous speech, with the possible exception of the fixed speaker names and corresponding spellings if these are available before the FDB database is recorded.

Utterance description	Items per call
Credit Card Numbers	1
PIN	1
Sequence of digits	1
Common sentences	4
Speaker-specific sentences	6
Spelled name / word	2
Speaker-specific application words	2
Speaker-specific names (from a set of 10)	2
Uttered speaker name (fixed)	1
Spelled speaker name (fixed)	1

Table 5.1: SpeechDatII SDB contents definition.

### 5.1 Digit/number strings

Digit strings have been used in speaker verification/identification for many years. Their origins can be traced back to audio text applications such as home banking where account numbers and their corresponding PINs are transmitted by DTMF. Keyed in DTMF numbers are susceptible to fraud as they can be used by an impostor. By contrast, verbal uttering of digit strings in a verification environment reduces the chances of fraudulent use significantly. Three types of digits/digit-strings have been adopted as follows.

#### 5.1.1 Credit card numbers

Each speaker will utter a 16 digit number once per call. A unique number will be assigned to each speaker, drawn from a list of 150 numbers, and this single number will be used in all 20 calls by that speaker. A list of the designed numbers is enclosed in appendix A.

The number is arranged in a credit card number format in blocks of 4 digits, i.e. XXXX XXXX XXXX XXXX or alternatively in blocks of 4, 6 and 5. The numbers are composed in a manner that will provide a fair tri-phone like distribution of digits as well as providing a realistic checksum as the last digit of the string. Details of how to compute the checksum is described in the FDB specification document (SD1.1.1). These credit card numbers are seeded in the Fixed Database (FDB) and the Mobile Database (MDB) for impostor data collection.

As these numbers will be read with no special prompting, they will probably contain both digits and free-form number expressions, such as “twenty-three hundred”. This is likely to be the case especially in the French database.

### **5.1.2 Personal Identity Numbers (PINs)**

Each speaker will utter a 6 digit PIN number once per call. A unique PIN will be assigned to each speaker, drawn from a list of 150 numbers, and this number will be used in all 20 calls by that speaker. A list of the designed 6-digit PINs is enclosed in appendix B.

The PINs are composed in a manner to cover as many tri-phone like combinations as possible. These PINs are also seeded in the FDB and MDB for impostor data collection. It is intended that the PINs are read as connected digits and not free-form numbers, also for these items in the FDB.

## **5.2 Sequence of digits**

This item is included to elicit isolated digit examples of all Arabic numerals for each speaker. The collected material will allow generation of impostor digit sequences for any digit based speaker verification system.

Each speaker will read the Arabic numerals (0 to 9) with sufficient pause in between numerals. Special instructions must be included in the prompt sheets to brief the speakers as to how they should read the digit sequence. The digit sequences should also be randomised to provide some variability in the position of the 10 numerals. It is expected that the recording duration will be much longer than for other items to avoid truncating any digits inadvertently.

## **5.3 Phonetically rich sentences**

Each speaker will utter two sets of (phonetically rich) sentences in each of the 20 calls. In the context of speaker verification, a phonetically rich sentence is defined to be the one with an adequate number of language-specific phone combinations which provide maximum speaker-specific information.

### **5.3.1 Common sentences**

A set of 4 different sentences common to all calls for a speaker but varying over the set of speakers are drawn from a pool of 250 sentences. Each speaker is expected to utter the same four common sentences in all his/her calls during the collection exercise.

In order to produce a corpus of impostors for the common sentences, it is recommended that an extra item from the SDB set of common sentences are included in the FDB collection.

### **5.3.2 Speaker-specific sentences**

A set of 6 sentences specific to each of the 20 calls of each speaker is defined. The 120 sentences for a speaker are selected from 5 separate sets of 120 sentences (from a total pool of 600 different sentences, distinct from the 250 common sentences). Care must be taken to ensure a good overall distribution of context independent phones when designing the sentences.

## **5.4 Spelled name / word**

Each speaker will spell 2 unique names or words at each call, that is a total of 40 different spelled words or names (the names spelled by each speaker may be drawn from the same set, but no name is repeated by a speaker in any of their calls). Assuming an average length of 7 letters per word, there will be 280 letters per speaker. It is important that a balanced repetition of letters is achieved for each language and especially for each speaker; for the case

of British English it is expected to be approximately 10 examples of each letter for each speaker.

### 5.5 Speaker-specific application words

Each speaker is assigned two specific application words selected from a list of 30 as described in the FDB specification document SD1.1.1. These two words are uttered in each call made by a speaker. All 30 application words should be utilised (approximately equally).

### 5.6 Speaker-specific names

A set of 10 typical “forename\_surname” combinations for each language is to be drawn for each speaker from the 150 names as defined in the FDB specification document SD1.1.1. Each of the 150 names should appear approximately uniformly. These items are designed specially to provide training and testing data for speaker-dependent name-dialling experiments, and to cover different recording conditions and handsets.

Each speaker is required to provide two (different) names from this list in each call. Of the 40 items thus provided by a speaker, each name should appear 4 times, 2 for each of GSM and fixed-network conditions, and ideally uniformly over the noise condition. Given the 70%/30% split in quiet/noisy call conditions it is not possible to have all names recorded in all 4 different acoustic conditions, however it is possible to prearrange the prompt sheets so that the same conditions are utilised by many of the name lists, which will simplify later experiments and analyses.

### 5.7 Uttered and Spelled speaker name

Each speaker is assigned a unique name (in the forename surname format) from a set of 150 names as defined in the FDB/SDB contents specifications. The assigned name will be uttered by the speaker at each call. The same name is also spelled verbally as a separate item.

## 6. Recording speaker-specific details

Recording speaker-specific details are of prime importance for any speaker verification database. Such details will allow researchers to form a more accurate understanding of possible factors contributing to the success or failure of speaker verification algorithms. However, one must be careful with the public’s sensibilities when raising questions concerning an individual’s personal details. For example, many women will find it very difficult to give an indication of their exact age. To avoid any possible misleading answers, it is wise to phrase the questions broadly. For instance, speakers should be asked to indicate their age bracket and not their exact age.

Attributes to be recorded for each speaker are as follows:

- Sex: male, female
- Age bracket is as described in SD1.2.3 and quoted below.

Age	Middle point	Recommended minimum %
0 ...15	8	recommended
16 ... 30	23	20
31 ... 45	38	20
46 ... 60	53	15
61 ... ~	78	optional

- State of health at the time of making each call: cold, sore throat, physically/mentally exhausted, fine
- Life style: smoker, non-smoker (relevant questions are included in the registration sheet and not as a part of the recording session).
- Subjective assessment of noise by speakers; e.g. quiet, loud, etc.
- Handset information.
- The regional accent of the speaker.

DTMF tones may be used to elicit all the relevant attributes, which would speed up the process of validation and annotation.

## **7. Recording environment**

As well as recording speaker-specific information, it is equally important to register details of the recording environment for each call. Two distinct environments have been envisaged for the purpose of data collection. They are home/office and public places such as streets and train stations. Each recording session must be characterised by the type of environment the call is being made from. This information will be elicited for each call. At the validation stage, Signal-to-Noise Ratio (SNR) of speech files are computed and registered as additional information. Although there are various definitions of SNR; it was decided to adopt the NIST's definition for the purpose of the SDB. Further details on the algorithmic aspects of NIST's SNR may be found in appendix C.

Each speaker makes half of his/her calls using mobile phones. 70% of the mobile calls are made from home/office (i.e. 7 calls) and the other 30% (i.e. 3 calls) from public places. The other 10 calls are made using terrestrial lines; out of which 70% is made from home/office (i.e. 7 calls) and 30% from public places (i.e. 3 calls).

## **8. Speaker recruitment**

Close biological relatives of a speaker are natural candidates to make the database more realistic as they are most likely to provide very similar voice prints to that of the target speaker. This is particularly of research interest if the biological relative is a twin. All partners are encouraged to make use of close biological relations when recruiting speakers; but, this is not a mandatory requirement as there may be unforeseen logistical problems.

## **9. Utterance elicitation and prompt sheets**

The FDB document SD1.1.1 describes different ways of utterance elicitation and designing prompt sheets which partners are referred to. Nevertheless, there are a number of issues to bear in mind when designing the collection exercise. They are summarised as follows:

- The utterances must be collected by reading a prompt sheet.
- As the speakers have to make 20 calls in total, they must be issued with a unique prompt sheet and ID number for each call.
- Although all calls may be completed within 2 months, partners must be aware of the fact that not all speakers will be well disciplined to observe the original time table. A period of 3-4 months is more realistic, considering the minimum interval required for each call.
- As it requires a high level of perseverance on the part of speakers, they must be thoroughly briefed and motivated by whatever means possible prior to commencing the data collection.

- The instructions must be clear. They must be written in a way that leaves no ambiguity as what the speaker should do. It is worth noting that overwhelming speakers with too many instructions will create enormous problems and are to be avoided at all costs.

## 10. Appendix A

The following is a set of 150, 16 digit credit card numbers generated by a computer program. Care was taken to ensure a fair tri-phone like distribution of digits. The last digit in the string is the checksum which was computed using a formula believed to be used by credit card companies. Details of the checksum algorithm may be found in LE2-400-SD1.1.1.

0500 8824 0710 2777	2557 1781 1365 7073
0490 2332 8674 1248	2442 8369 8616 5198
0510 0620 7667 3063	2562 1665 7368 2262
0480 8636 0800 9882	2432 0644 8695 8697
0520 2592 0833 4444	2572 0165 1591 0008
0470 1191 0364 9173	3503 3393 9315 6763
0530 9627 0888 9258	3493 1581 8668 9356
0460 6894 1601 2055	3513 3373 0400 6984
0540 2622 7895 8262	3483 1331 0764 7082
0455 3633 2402 8994	3523 3334 8802 8983
0555 0744 1721 0154	3473 9607 8626 0064
0440 1681 6295 4013	3533 0610 8714 6744
0560 0773 8387 0250	3463 2612 6195 0283
0430 9686 2602 6254	3543 6385 7875 5345
0570 3367 1735 1512	3458 0264 0590 7268
1501 0330 6973 4096	3558 8324 9325 7998
1491 1765 7405 2082	3443 1135 8304 2171
1511 2712 1274 1075	3563 7673 0320 9560
1481 6604 8812 7025	3433 9963 0234 1029
1521 3892 1201 5817	3573 3383 0220 7985
1471 0873 9933 1177	4504 0390 1291 7853
1531 1611 2666 7589	4494 9796 0420 0055
1461 2722 0290 7171	4514 1311 2322 9472
1541 8314 6395 9249	4484 7196 1211 9026
1456 9912 9725 2063	4524 0380 0110 9833
1556 8287 3613 9650	4474 2312 1301 6991
1441 2682 2366 3160	4534 2372 0664 4630
1561 2672 3583 9796	4464 8876 0680 9870
1431 1371 7635 1754	4544 0280 7396 7065
1571 9779 4095 9865	4459 7097 0823 0176
2502 0890 0300 6881	4394 0410 9705 2153
2492 2632 6096 2266	4384 1791 8685 1713
2512 7386 4186 5995	4294 2292 9923 2043
2482 1671 9896 1997	4284 1691 7296 9059
2522 6624 2093 1166	4194 9298 6775 0026
2472 0600 0864 3990	5505 8098 0844 1052
2532 3623 3185 3140	5495 8834 2184 6460
2462 1183 1811 9853	5515 1391 6594 8152
2542 8397 6634 5138	5485 2302 2692 8220
2457 3423 0670 0089	5525 9835 7188 4156

5475 6874 7286 9048	7585 9388 0900 9086
5535 1381 7615 9078	7415 7323 9617 9030
5465 8224 8297 9188	7595 7713 8704 9267
5545 0273 1421 8169	8508 3403 8855 2002
5683 8795 6614 0075	8498 9203 8189 9849
5693 3293 2392 7677	8518 9786 0182 6089
5783 2275 7776 0144	8488 6285 1411 3077
5793 1701 8278 9454	8528 7703 7684 3680
5889 1281 0370 3264	8478 2412 9825 4083
5893 9803 8865 3174	8538 2422 9669 3890
6506 9675 8376 9555	8468 6187 9288 3081
6496 8724 0754 7884	8566 0210 7625 9584
6516 7335 1235 2121	8436 2192 8778 3081
6486 6404 0034 1080	8576 3593 9099 9164
6526 1631 1265 8057	8426 1774 0124 3469
6476 8606 3276 2992	8586 1621 0580 0836
6536 8845 7277 8261	8416 1114 0854 9315
6466 0700 0310 8146	8596 9377 3413 7579
6559 7694 0790 8062	9509 2891 9715 5985
6564 0720 9225 9062	9499 3283 9213 9360
6434 1035 7303 3557	9519 9115 0810 0016
6574 9398 1755 7160	9489 9101 9813 4943
6424 8785 1645 1048	9529 0630 0091 0043
6584 7723 7974 9158	9479 1092 6374 8172
6414 8202 3193 8083	9539 3094 8212 9133
7507 9944 9407 5358	9469 0200 2702 1263
7497 3792 1801 9340	9567 9279 9901 9649
7517 3782 0780 8040	9437 3603 0690 6068
7487 7313 2382 2079	9577 1711 2223 7376
7527 9305 1745 7807	9427 9954 0734 7654
7477 1125 7605 2163	9587 2282 8406 1065
7537 0100 1321 8660	9417 6784 1221 8076
7467 6794 0134 9982	9597 8975 9976 9998
7565 2582 9696 2141	
7435 7375 1401 9147	
7575 8197 9198 9174	
7425 0654 0190 0241	

## 11. Appendix B

The following is a set of 150, 6 PINs generated by a computer program. Care was taken to ensure a fair tri-phone like distribution of digits.

000100	122123
002003	127128
004005	129132
006007	133134
011012	137138
013014	139140
015016	141420
019021	143144
022023	147148
024025	149152
026027	153154
030310	155156
032033	159160
034035	161621
036037	163164
041042	167168
043044	169172
045046	173174
049050	175176
051052	181820
053054	183184
057058	185186
059061	193194
062063	195196
066067	199201
068069	222300
070710	224211
072073	226227
076077	229214
081082	236237
085086	248249
087088	255256
089090	257217
091092	265266
095096	272730
097098	283281
099111	284285
112102	293290
115116	296297
117118	298299
119120	333411

336302	567524
337322	568562
338339	569563
347348	576536
354352	577578
355356	579580
364365	586581
366367	587583
374370	596592
378379	666700
384381	668605
389390	669612
395393	676613
398399	677614
444540	678615
447415	679662
449404	687673
456422	688633
457441	689681
466443	697682
467460	698690
469463	699722
475472	777811
485480	779733
486481	787702
488489	788723
495492	789745
496497	798785
498499	799801
555601	888922
557500	898804
559523	999800
565611	958858
566502	907807

## 12. Appendix C

There are different definitions of SNR but for SDB purposes it is taken to be

$$SNR = 10 \log_{10}(\text{peak\_signal\_power} / \text{mean\_noise\_power})$$

A program called “stnr” is available by anonymous ftp from the NIST’s ftp site ([jaguar.ncsl.nist.gov](ftp://jaguar.ncsl.nist.gov)). It is part of spqa-2.3 package which may be found under “*pub/spq\_2.3+sphere\_2.5.tar.Z*”.

The procedure for computing SNR is described in the NIST’s accompanying document; however, its main features are briefly outlined in the following:

The power refers to the variance of the speech signal computed over a 20 ms window. A signal energy histogram is generated by computing the root mean squared (RMS) power, in decibels, over a 20 ms window and then updating the appropriate histogram bin. The window is then shifted by 10 ms and the next power is computed.

At high SNRs the histogram will consist of two peaks; the lower one originating from the noise and the higher from the signal plus noise. At low SNRs there will be overlap of these peaks. A raised cosine function is fitted, in the Chi-squared sense, to the left hand peak of the complete RMS histogram in order to obtain an estimate of the noise power distribution. The mid point of the raised cosine is labelled as the mean noise power level. This cosine function is then subtracted from the histogram to obtain the speech power distribution. The speech level is defined to be the histogram bin midpoint where the 95th percentile occurs in the speech power histogram.