

**DELIVERABLE IDENTIFICATION**

Identification number	LRE-63314-D1.3.1&D1.3.2
Type	Technical Report
Title	Specification of Short/Mid-term Databases
Status	Final Version
Deliverable	D1.3.1 & D1.3.2
Work Package	WP 1
Task	Task 1.3
Period covered	March - August 1995
Date	Oct 1995
Version	Final
Number of pages	14
Author(s)	David Pearce
Work package (WP) / Task (T) responsible	David Pearce (GEC) / David Pearce (GEC)
Project contact point	Harald Höge, 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	José Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	
Key words	
Abstract	
Status of abstract	

Received on	
Recipients catalogue number	

**DOCUMENT EVOLUTION**

Version	Date	Status	Notes

SPEECHDAT

WORKPACKAGE 1.3

DELIVERABLE D1.3.1 and D1.3.2

editor D J B Pearce

Version: Final

**Specification of Short/Mid-term Databases**

Oct 1995

CONTENTS

1. INTRODUCTION
2. FORMAT OF MODEL TEMPLATE FOR DATABASE OPPORTUNITY PROPOSALS
3. DATABASE SPECIFICATIONS
  - 3.1 Polyphone Telephone Database
  - 3.2 Mobile Telephone Database
  - 3.3 Spontaneous Telephone Application Database
  - 3.4 Automobile Database
  - 3.5 Telephone Speaker Verification Database

## 1 INTRODUCTION

The objective of workpackage 1.3 were:

- 1) To develop a model framework for the specification of speech databases.
- 2) To agree and specify the short/mid-term speech database needs using this model.

This deliverable report documents the specification of an agreed and prioritised set of databases needed in the short to mid-term within Europe.

For this exercise the timeframes were defined as:

- short-term                    1-2 years
- mid-term                      2-4 years.

It should be noted that these databases will continue to have utility well beyond these time frames. The timescales were really an issue of priorities addressing the question of what databases would be of the most utility to support current technology and pull through further R&D developments?

The databases presented here represent the consensus view of the partners contributing to workpackage 1 of the Speechdat project, "Short and Mid-term Databases for Applications". These partners are CSELT (Italy), IDIAP (Switzerland), GEC-Marconi (U.K.), JYDSK (Denmark), Philips (Germany), Portugal Telecom/INESC (Portugal), Siemens (Germany), SPEX (Netherlands), Vocalis (U.K.). The databases selected and their design reflects the needs identified within Workpackage 1.1 on the "Database Requirements for Teleservices".

The process used to define the databases and reach consensus agreement has been as follows:

- 1) A model template was developed for the specification of databases (presented in section 2 of this report)
- 2) Collection of proposals from Speechdat partners by e-mail.
- 3) Sorting of proposals and agreeing a shortlist of priorities (Resulting from discussions held at the Lisbon Speechdat meeting, 22nd May)
- 4) Refinement and merging of similar proposals for the specification of the selected shortlist of databases.
- 5) Review and feedback of the draft of this report.

Five different databases have been identified and are presented in section 3. Of these databases the priority need is for the following three databases:

- 1 Polyphone Telephone Database
- 2 Mobile Telephone Database
- 3 Telephone Speaker Verification Database

## **2 FORMAT OF MODEL TEMPLATE FOR DATABASE OPPORTUNITY PROPOSALS**

(this version gives instructions on what information to supply)

### **ONE PAGE DATABASE OPPORTUNITY PROPOSAL**

Title:

Author:

Organisation:

Date:

**DATABASE DESCRIPTION / TARGET APPLICATION:(6 lines max)**

This field contains a short description of the purpose for which the database has been proposed, its importance in term of scientific knowledge and/or the economic relevance of the target application.

**SPEECH PROCESSING DEVICE NEEDS:(6 lines max)**

The main characteristics of the speech processing device (ASR, TTS, Coder ...) for which the material is collected, are described.

**SPEAKER/TALKER POPULATION:(6 lines max)**

This field will contain a broad definition of the talker population (characteristics, number,...)

**SIGNAL CHARACTERISTICS:(4 lines max)**

The signal characteristics as environment, microphone, transmission channel are listed.

**VOCABULARY AND TYPE OF SPEECH:(6 lines max)**

The vocabulary is described and the method of speech elicitation are mentioned (read speech, spontaneous ....)

**MULTILINGUAL OPPORTUNITY:(4 lines max)**

The languages used are listed here and the added value in having a multilingual material, if any, is explained.

**DRAFT EVALUATION OF DATABASE SIZE:(4 lines max)**

A gross evaluation of the size of the db is provided in MBYTES, taking into account the vocabulary size, # speaker, type of coding.

**TYPE OF PROCESSING:(6 lines max)**

In this field, the step of processing needed to the speech material are listed (verification, word labelling, phonetic labelling ...)

**COST ESTIMATION:(4 lines max)**

A gross estimate of the cost of the speech database is provided taking into account the processing claimed for the speech material

**ALTERNATIVES:(6 lines max)**

Possible alternatives, already available or not, are listed to allow an objective evaluation of the proposal

### 3 DATABASE SPECIFICATIONS

#### 3.1 Polyphone Telephone Database

##### DATABASE DESCRIPTION / TARGET APPLICATION:

The database is targeted at the development of speech recognition capabilities to support teleservices accessed by voice over the telephone network.

Telecommunications is the largest market for speech recognition. There are therefore economic benefits since these databases will permit the development of these services.

They also provide material for research into improved algorithms, and investigation into the differences between languages and accents.

##### SPEECH PROCESSING DEVICE NEEDS:

The database will allow the training and testing of speaker-independent recognition for whole-word and sub-word type recognisers. In addition, the database can be utilized for speaker-independent spotting of command words embedded in extraneous speech. While the database is suited to the training of statistical recognisers based on HMMs it is also technology independent.

##### SPEAKER/TALKER POPULATION:

5000 speakers.

Population sampled to give representative coverage of the range of voices by gender, region (accent) and age.

##### SIGNAL CHARACTERISTICS:

Collected over the national terrestrial telephone network. 8kHz sampling, A-law companding. Standard telephone handsets including cordless phones to be used. Database recorded by an automatic recording platform with a digital connection to the telephone network.

##### VOCABULARY AND TYPE OF SPEECH:

Vocabulary to consist of the following parts which will be distributed across the speaker population:

Isolated common application words, connected digits, natural numbers, money amounts, dates, times, spelled names, "yes"/"no", and sentences giving phonetic coverage.

Read speech from prompt sheet + spontaneous responses to questions.

##### MULTILINGUAL OPPORTUNITY:

All European languages including language variants (e.g. French as spoken in Belgium and Luxembourg, in addition to French as spoken in France).

Benefits result from being able to provide multi-lingual applications.

##### DRAFT EVALUATION OF DATABASE SIZE:

8 GBytes per language.

##### TYPE OF PROCESSING:

Orthographic transcription of each file.  
Validation by human listener that correct material was spoken.

COST ESTIMATION:

200KECU per language.

ALTERNATIVES:

## 3.2 Mobile Telephone Database

### DATABASE DESCRIPTION / TARGET APPLICATION:

The purpose of the database is for the testing and training of recognisers over mobile telephone networks. The content will be similar to the Polyphone but collected over mobile telephone networks with particular emphasis on the full-rate GSM network. The database enables the influence of background noise and channel on the ASR performances to be evaluated and technology and algorithms improved.

The number of mobile phone users in Europe is growing significantly. Mobile users will also want to access teleservices.

The database supports research into noise and channel robust algorithms for speech recognition.

### SPEECH PROCESSING DEVICE NEEDS:

Testing ASR of isolated/continuous speech in telephone mobile environment:

- a) Testing recogniser which has been trained on a non-mobile telephone database (e.g. Polyphone)
- b) Testing recogniser which has been trained on mobile data. This database needs to provide both the training and testing material for training and testing.

### SPEAKER/TALKER POPULATION:

1000 speakers.

Population sampled to give representative coverage of the range of voices by gender, region (accent) and age.

### SIGNAL CHARACTERISTICS:

The data will be collected with good connections over a range of real-world background noises e.g. quiet office, public place, pedestrian by road side, in a stationary car (windows shut), as a passenger in a moving car.

Recordings will be made of A-law PCM transcoded speech originating from mobile networks. The collection platform will have a digital connection to the terrestrial network.

Only speech from mobile handsets will be collected (i.e. not hands-free)

### VOCABULARY AND TYPE OF SPEECH:

Polyphone like vocabulary but with some modification to ensure sufficient numbers of occurrences of items (over the range of background noise conditions) for testing. Some parts of the text material should be identical with the polyphone recordings to enable comparison of recogniser performance on mobile data if training has been carried out either on mobile data or on data recorded from the terrestrial network.

See Polyphone telephone database specification.

### MULTILINGUAL OPPORTUNITY:

If noise and channel robust algorithms are developed then systems can be trained on the Polyphone databases and will give adequate performance for mobile users. In this case the

database needs would be language-independent. If, however, training on mobile data gives significant performance improvement then multilingual databases are needed. It is therefore recommended that mobile databases be collected for a few languages first and only expended to further languages if experiences show this to be necessary.

**DRAFT EVALUATION OF DATABASE SIZE:**

2 GBytes per language

**TYPE OF PROCESSING:**

Orthographic transcription of each file.

Validation by human listener that correct material was spoken.

It is not necessary to systematically control or transcribe channel effects such as fading or bit errors.

**COST ESTIMATION:**

70KECU

**ALTERNATIVES:**

### 3.3 Spontaneous Telephone Application Database

#### DATABASE DESCRIPTION / TARGET APPLICATION:

Spontaneous speech collected when users are interacting with an automatic service based on current recognition capabilities (i.e. isolated word or word spotting of small active vocabularies). The purpose of the material is as a test database to evaluate and improve the recognition technology to cope with real-world data.

#### SPEECH PROCESSING DEVICE NEEDS:

The database will support the testing of speaker-independent automatic speech recognition systems. In particular it will test how recognisers handle the following: - spontaneously spoken utterances, out-of-vocabulary utterances and embedded phrases requiring key-word spotting.

#### SPEAKER / TALKER POPULATION:

500 to 1000 native speakers.

#### SIGNAL CHARACTERISTICS:

Collected over the national terrestrial telephone network. 8kHz sampling, A-law companding. Standard telephone handsets to be used. Database recorded by an automatic recording platform with a digital connection to the telephone network while interacting with a service based on small vocabulary recognition.

#### VOCABULARY AND TYPE OF SPEECH:

Service based on a small number of active words (<100).  
Spontaneous interaction with telephone application.

#### LANGUAGE / MULTILINGUAL OPPORTUNITY:

All European Languages.  
The database serves as a testing benchmark for each language. It enables the performance of a recogniser trained on polyphone material to be tested on real-application data.

#### DRAFT EVALUATION OF DATABASE SIZE:

2 MBytes per speaker.

#### TYPE OF PROCESSING:

Transcription of spoken input by human listener.  
Categorised by type e.g. in-vocabulary, out-of-vocabulary, embedded etc.

#### COST ESTIMATION:

60 KECU per language.

#### ALTERNATIVES:

### 3.4 Automobile Database

#### DATABASE DESCRIPTION / TARGET APPLICATION:

The database is targeted at the development of speech recognition capabilities to support speech driven applications in the car environment. Typically, there are not only time varying acoustic conditions but also an additional speech variability caused by traffic situation and noise environment (LOMBARD effect).

The economic relevance lies in applications of speech technology in the car and improving safety. Now emerging traffic guidance systems and other devices need robust recognition systems for large vocabularies.

The database provides material for research into robust recognisers and noise reduction algorithms.

#### SPEECH PROCESSING DEVICE NEEDS:

The database will allow the training and testing of speaker-independent recognition for sub-word type recognisers. It will also allow investigation of the mismatch between speaker-dependent training in a stationary vehicle and testing when driving. While the database is suited to the training of statistical recognisers based on HMMs it is also technology independent.

#### SPEAKER/TALKER POPULATION:

500 speakers.

Population sampled to give representative coverage of the range of voices by gender, region (accent) and age.

Emphasis should also be placed on coverage of different types of car environment and driving conditions.

#### SIGNAL CHARACTERISTICS:

Collected over PC-based devices or DAT-Recorders.

8 or 16kHz sampling, A-law companded or 16 bit linear. The utterances should be made in a parked vehicle and during real life driving conditions. Database recorded by an human supervised platform with different types and a number of microphones including a hands-free microphone.

#### VOCABULARY AND TYPE OF SPEECH:

Vocabulary to consist of the following parts which will be distributed across the speaker population:

- connected digits, natural numbers, place names, person names, spelled names and places.
- application-independent: phonetic sentences
- application-dependent words: embedded commands for car stereo operations, phone operations and traffic guidance

Utterances should be prompted as well as spontaneously elicited through loudspeaker in the car.

#### MULTILINGUAL OPPORTUNITY:

All European languages are needed.

DRAFT EVALUATION OF DATABASE SIZE:

1 GigaByte per language.

TYPE OF PROCESSING:

Book keeping of driving conditions

Orthographic transcription of each file.

Validation by human listener that correct material was spoken.

COST ESTIMATION:

140 KECU per language.

ALTERNATIVES:

- Mixing of the car environment noise with previously recorded speech.
- Acoustic simulation of the car environment during the speech recordings.

### **3.5 Telephone Speaker Verification Database**

#### **DATABASE DESCRIPTION / TARGET APPLICATION:**

The purpose of the database is for the development and assessment of speaker verification systems for security. The database needs to capture the intra speaker variability over time.

The economic drivers are to support teleservices requiring security (e.g. for telebanking) and to reduce telephone fraud (estimated at \$3.3 Billion in U.S. in 1994). The database would allow parameter optimisation and performance assessment of verification systems.

A verification database will enable research into improved speaker recognition algorithms.

#### **SPEECH PROCESSING DEVICE NEEDS:**

The database will be used for training and evaluating speaker verification systems. It shall contain material for training and testing verification systems that are based on a restricted vocabulary, as well as those operating on a flexible vocabulary or using text-independent approaches.

#### **SPEAKER/TALKER POPULATION:**

40 speakers, each one recorded 60 times.

Since short term intra speaker variability is not expected to depend on one's age, education or social status, these sample criteria may be omitted. An equal number of male and female speakers should be recorded.

Speakers should be of the same regional accent.

#### **SIGNAL CHARACTERISTICS:**

Collected over the telephone network. 8kHz sampling, A-law companding. Standard telephone handsets to be used; each speaker to use a small number of different handsets. Database recorded by an automatic recording platform with a digital connection to the telephone network.

#### **VOCABULARY AND TYPE OF SPEECH:**

Utterances should be a combination of read, spontaneous material and audio prompted. The content will be designed for a variety of realistic operational scenarios and permit training and testing. The vocabulary would include: isolated keywords, isolated digits, connected digits and phrases.

The recording sessions for each speaker should be distributed over a period of at least 2 months and include a variation of individual speaking conditions (time of day, emotion, state of health).

#### **MULTILINGUAL OPPORTUNITY:**

It is expected that speaker verification algorithms should be largely language-independent.

Databases for a small number of European languages would enable this to be tested.

#### **DRAFT EVALUATION OF DATABASE SIZE:**

5 GBytes per language.

#### **TYPE OF PROCESSING:**

Orthographic transcription of each file.  
Validation by human listener that correct material was spoken.

COST ESTIMATION:

80 KECU per language

ALTERNATIVES: