Catalan FreeSpeech Database

1. BASIC INFORMATION

1.1. Resource description (broad description of the database, language)

The FreeSpeech Catalan database was recorded in 1999 for automatic dictation purposes. 148 speakers (100 adult, 48 children) belonging to the 4 main Catalan dialects read texts from several domains. The signals were recorded at 16 kHz with a headset microphone and a FreeSpeech (IBM) mouse, which has a built-in microphone. The Catalan FreeSpeech Database was funded by the Catalan Government.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Javier Hernando
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016433
Fax: +34 93 401 6447
e-mail: javier.hernando@upc.edu

2.2 . Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive.

2.3 . Copyright statement and information on IPR

The resource has copyright. The Copyright belongs to Universitat Politècnica de Catalunya. The resource is free, license-based, for research purposes and fee license-based for commercial purposes.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain several files containing documentation (copyright, readme), speech files (one per utterance) and label files (each speech file has an accompanying label file) in a nested structured directory.

3.2. Encoding

Documentation is encoded in plain text.

Speech files are stored as sequences of 16-bit 16 kHz uncompressed speech samples. Each prompted utterance is stored within a separate file.

Each speech file has an accompanying XML/ASCII SAM label file Label files contain information about the database, speech signal coding, speakers, turns, segmentation, labeling session, transcriptions and annotations.

3.3. Size of the resource (size of recorded speech/MB occupied on disk)

The corpus contains about 120 hours of recorded speech (about 15 GB)

4. CONTENT INFORMATION

4.1 Type of the resource (language, ASR, BN, dialogues, TTS….)

This is a database for training ASR systems in Catalan for dictation applications.

4.2. Content description. This section explains the basic content of the data: Corpus design (read, spontaneous), recorded data (BN, parliament), recorded dialogues,…

The text read by the speakers was provided by Philips, and includes a wide diversity of topics and writing styles.

4.3. Transcriptions, annotations: Types of annotations (orthographic, phonetic, noises, pitch, breaks, segmentation…), Reliability of the annotations

Only the texts that were read by the speakers are provided. The recordings were not posteriorly checked for consistency with the read text.

4.4. Lexicon. Description of the lexicon (if applicable)

N.A.

4.5. Speakers. Number of speakers, recruitment strategies, accents/ regions, age, gender

The database contains recordings from 148: 50 women, 50 men and 48 children. Speakers were selected from the four main Catalan dialects.

4.6. Recording platform

4.6.1.Domain(s), environments,

Recordings were performed in a sound proof room

4.6.2 Recording platform

The recording platform is based on equipment provided by Philips. The signals were recorded at 16 kHz with a headset microphone and a FreeSpeech (IBM) mouse, which has a built-in microphone.