AGORA

1. BASIC INFORMATION

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

The Agora database contains the recordings of 34 TV shows of Catalan public broadcast TV3. The shows are highly moderated debates with a high variation in topics and invited speakers. The database consists of 68 files with a total audio time of 43h. Each file corresponds to half show of an airing day with an average duration of 38 min. The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events. The transcriptions have four layers. Transcriptions follow the TRS format produced by the Transcriber transcribing tool.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: José A. R. Fonollosa
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 4016439
Fax: +34 93 401 6447
e-mail: jose.fonollosa@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

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, documents, …), video and audio files, and transcription files in the same directory.

3.2. Encoding

Documentation is encoded in word/pdf/plain text.

64 video are stored as MPEG-2. Their coresponding audio files are stored as 16-bit 32 kHz uncompressed speech samples . Files have an average duration of 38 min.

Each audio file has an accompanying XML transcription file.
The XML transcription files contain information about the database, speakers, turns, segmentation, background sounds, channel and literal transcriptions.

3.3. Resource size (size of recorded speech/MB occupied on disk)

The corpus contains about 43 hours of recorded speech.

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 broadcast news and TV debates 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 database contains the recordings of 34 TV shows of Catalan public broadcast TV3

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

The transcription follows the general guideline generated within the TC-STAR project for European Parliament Plenary Sessions but it was extended to include additional information as the language, background condition, silence/voice segmentation, speaker segmentation and acoustic events. The transcriptions have four layers. Transcriptions follow the TRS format produced by the Transcriber transcribing tool.

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

Not applicable

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

The database recordings contain segments from 871 adult Catalan speakers (441 male, 113 female, 317 unknown), and 157 adult Spanish speakers (83 male, 29 female, 45 unknown). Speakers may originate from different accents. Speakers are unbalanced in gender favouring male speakers in total duration.

4.6. Recording platform

4.6.1.Domain(s), environments,

All the shows were performed in a closed TV studio

4.6.2. Recording platform
Not known