UPC ESMA

1. BASIC INFORMATION

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

This database contains the recordings and annotations of read text material in neutral style. The database was recorded by one female Spanish professional speaker.

Speech was recorded in a noise-reduced room. The signal was recorded at 32kHz and 16 bits and decimated to 16kHz. It includes a second channel with the laryngograph signal. The speaker read text material in neutral style. The text material is composed by 506 phonetically balanced sentences, 208 phonetically balanced short paragraphs and 62 long paragraphs, giving a total of 1h 45min recorded speech. The database includes the phonetic transcription, phonetic segmentation and epoch labels.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Antonio Bonafonte
Address: Jordi Girona 1-3 Edifici D5, 08034 Barcelona, Spain
Affiliation: TALP research center. Universitat Politècnica de Catalunya
Position: Professor
Telephone: +34 93 401 0764
Fax: +34 93 401 6447
e-mail: antonio.bonafonte@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, phoneme statistics), speech files (one per utterance) and manual and automatic label files including the phonetic transcription and segmentation. (each speech file has an accompanying label file).

3.2. Encoding

Documentation is encoded in plain text.

The database is made of three subcorpus: sentences (SE subcorpus, 30 minutes), paragraphs (PA subcorpus, 30 minutes) and literary paragraphs (PL, 45 minutes).

For each utterance, two WAV signals are provided, one for the speech waveform (.wav extension) and one for the laryngograph signal (.lwv extension). These signals are stored as sequences of 16-bit 16 kHz with the least significant byte first ("lohi" or Intel format) as (signed) integers. Each prompted utterance is stored in a separate file. Furthermore, for each utterance, several label files provide the phonetic transcription, phonetic segmentation and CGI (closure glottal instant) labeling. Two sets of transcriptions are provided: manual (supervised) and automatic.

Automatic transcription:
For each utterance the phonetic transcription is derived using Saga (UPC open source tool). HMM based forced-alignment is used to detect pauses and segment the speech files into phonemes.

For each utterance, the following files are provided:

- .txt: prompt text. It includes the tag "<S>" to indicate pauses.
- .pho: phonetic transcription in SAM-PA, including syllable and word bondaries.
- .seg: phonetic segmentation: start and ending time of each phone.
- .prb: probability assigned to each segmentation (useful to detect potencial segmentation problems).
- .cgi: instants of the glottal closure (and implicitely, voiced/unvoiced and pitch labeling).

Manual transcription:
The manual transcription includes, for each utterance, the same files than the automatic transcription, except the .prb. Therefore, the prompt text (.txt), phonetic transcription (.pho) and segmentation (.seg) and the cgi instants (.cgi). However, manual labels are not available for the whole database. In particular, there is no manual information for the PL subcorpus and for some utterances in the PA subcorpus.

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

The corpus contains about 776 files / 1h 45min of recorded speech and needs about 450MB for disk storage.

4. CONTENT INFORMATION

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

This is a database for training TTS systems in Spanish. The recordings were produced at the same time and by the same female speaker than the Interface Emotional Speech Database.

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

The corpus consist of three subcorpora.

- Subcorpus SE: 506 phonetically ballanced sentences designed to maximize the phonetic variability, taking into account the stress and position in the sentence.
- Subcorpus PA: 206 short paragraphs from news.
- Subcorpus PL: 62 long paragraphs. The speaker portrayed a bit the text of this subcorpus.

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

As already mentioned, phonetic transcription, phonetic segmentation, pauses and closure glottal instants (CGI) are labelled automatically for all the files; and manualy for the sentences and some of the short paragraphs.

The automatic phonetic segmentation was created using UPC tools. Phonetic segmentation uses HMM-based forced alignments. In the first step, the HMM toolkit finds the pauses and the pronunciation variants. In the second step, the phonetic segmentation is derived. The automatic CGI labelling used the Praat program.

Note that the manual transcription was generated several years before
the automatic transcription. There maybe small inconsistencies in the transcription.

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

The lexicon file is an alphabetically ordered list of distinct lexical items (i.e. splitting tokens at every space, essentially words in our case) which occur in the corpus with the corresponding pronunciation information. Each distinct word has a separate entry. As the lexicon is derived from the corpus it uses the same alphabetic encoding for special and accented characters as used in the transcriptions.

4.5. Speaker.

The database was recorded by a female actress, from Barcelona.

4.6. Recording platform

The database was recorded in a isolated room at Universitat Politècnica de Catalunya. The studio includes two isolated rooms, one for the speaker and other for the operators. The speaker recorded simulatenously in two channels: membrane microphone, and laringograph. The recording platform was developed at UPC. It consists of a computer program that allows to navigate through the selected prompts and controls the synchronous recordings (asio interface). The platform shows the recording levels, clipping, etc. The signals were recorded at 32kHz and with 16 bits per sample..