OGI 22 Language Speech Corpus

Ron Cole, Terry Lander, Jacques de Villiers, Yeshwant Muthusamy

Note: The OGI 22 Language Speech Corpus was conceptualized and implemented by Ron Cole and Jacques de Villiers at the Oregon Graduate Institute's Center for Spoken Language Research. The 22 language corpus, and the preceding 11 language corpus, were developed to support Yeshwant Muthusamy's thesis research. Mr. Muthusamy, a graduate student at OGI, was interested in developing algorithms to identify language automatically. Mr. Jacques de Villiers provided the infrastructure that enabled the data collection. was interested in Ms. Terry Lander worked with Dr. Beatrice Oshika to develop and implement the phonetic transcriptions for each language. The 11 language and 22 language corpora were, at the request of Mr. Curt Boyles of NSA, donated to the Department of Defense, which supported a three year "friendly competition," held at Johns Hopkins. The training and test sets designed by Dr. Muthusamy were used in the competition. During Ron Cole's time as Director of CSLU, all speech corpora were freely distributed to the research community.

General Description

The 22 Language corpus consists of telephone speech from 22 languages: Eastern Arabic, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, Vietnamese, and English. Unfortunately, French is not available. The corpus contains fixed vocabulary utterances (e.g. days of the week) as well as fluent continuous speech. We were expecting at least 300 callers in each language. Each utterance is verified by a native speaker to determine if the caller followed instructions when answering the prompts. Some of the calls in each language are transcribed orthographically.

Recording Details

All of the data in this corpus were collected over digital telephone lines. The digital data were recorded with the CSLU T1 digital data collection system. These files were sampled at 8 khz 8-bit and stored as ulaw files.

All of the wave files were converted to riff format with 16-bit linear coding.

Directory Structure

There are several top-level directories in this distribution: docs, labels, misc, speech, trans.

The speech directory contains the speech data files. Each speech filename has the following structure:

```
ww-xxxxx.yyy.wav
```

`ww` = language abbreviation

`xxxxx` = call number

`y` = utterance type code

For example:

```
EN-105.nlang.wav
```

This utterance is from the English speaker 105 and contains the answer to the question "What is your native language?".

As a participant proceeds through the data collection protocol, he is asked a series of questions. Each of the responses is stored as a separate speechfile. The utterance type code relates the recorded utterance to the protocol questions. The description of the protocol shows all of the utterance codes.

These audio and text files are subdivided into directories based on their call number mod 10. So, these files would be found in `/speech/10`.

Verification

Each utterance included in the 22 Language Corpus has gone through a process of verification. Native speakers of each language did verification. The verifiers were asked to listen to each utterance and decide if the speaker responded appropriately to the prompt. In addition, the verifiers made judgements about the age, gender, and dialect of each speaker.

Two native talkers verified the utterances in each language independently. Subsequently, they reexamined each utterance for which there was disagreement and produced an info file containing the 'resolved' judgements. Note: we resolved differences in Spanish, Vietnamese and Swahili by choosing the person with the overwhelmingly correct responses. For the other languages in the corpus we resolved every disagreement by hand.

Initially we asked the verifiers to make two judgement that are not now included in the release:

- whether or not the speech was cutoff at either end of the utterance
- whether or not there was missing information in the file. Because these judgement were

unreliable, the information regarding cut off speech and missing information is not included in the distribution.