Open Eval KW Search Telecon 03 13 2013

Attendees: Jonathan Fiscus, David Joy, Martial Michel, Mary Harper, Ronnie Silber, Doroteo Toledano, Dimitra Vergyri, Nancy Chen, Evelyne Tzoukermann, Sunil, (someone we missed)

Agenda Items:

Questions:

Is it possible that s KW will be OOV

Evaluation KWs will be In-Test and Out-of-Test KWs as well as Out-of-Training KWs.

For less experienced team , are there any data sets that we should start with.

LDC datasets (e.g., the STD '06 data Mandarin and Arabic conversational data). The surprise language will be a non-English language.

What are the sizes of the data sets?

80 hrs of conversation training, ~20 hrs of scripted non-spontaneous training, 10 hrs conversational development data, and 75 hrs of conversational eval.

Look at specs for the data; also the LSP will give you info about language including phoneme set. You might also look at the Babel BAA section 1.A.1 (https://www.fbo.gov/index?s=opportunity&mode=form&tab=core&id=78fece5f8ad57ddffb437dff607446fc&_cview=1).

Lexicon Syllable Boundaries and Stress Markings

Pronunciation lexicon contains words in orthography, pronunciations in XSAMPA, along with Syllable boundaries and stress markings, when appropriate.

Segments and Time marking

There are a couple levels of time marks:

1. Transcriptions have segment-splitting times. They are not tightly end pointed and include silence so you're better off doing re-segmentation.
2. We will provide forced word alignments for the dev data. We use the forced alignment to remove non-speech time from the segmentation. This data will be in the *.mitllfa3.rttm and *std06fa.rttm files in the IndusDB releases.
3. NIST will use the same forced alignment process on the eval data but will not release them.

Aside: The scripted data is an in addition to the 80 hrs of transcribed conversational speech. There are some phonetically balanced sentences-which might help. Start with the Babel Data Spec to understand the data and how it is organized.

Performance metric question: which metric is primary ATWV or WER?

ATWV is the current primary metric of the Babel program. If you're doing it for you own purposes you can use what you want. You may want to optimize differently depending on whether you care more about WER or ATWV.

Will there be extra text for language modeling data?

No. We don't provide extra text for Language Model. You could go to web, but keep in mind that using extra text outside of Babel, changes the category of submission to "OtherLR".

Eval Plan and Speech-Nonspeech Detection

Eval Plan Section 3.1- If you use your homegrown speech non speech detector trained on non-Babel language, this will an "OtherLR" system. OpenEval participants can choose to submit what ever condition they wish to explore.   The Babel performers must submit a baseline model trained only on the surprise language build pack.

Make sure you describe your system well enough in the system description and be sure to pick the right category via your Experiment ID.   Look at the Spec and Eval plan and what other resources you are using. In terms of comparing systems, it's important to know the category of submission.

NIST compares systems for like conditions. We give people the ability to submit contrastive runs that are used for within-team comparisons.

We plan to release some surprise language KWs with the build pack so that you can play around with KWS in dev while you're building your final system.

Install-and SSH Keys

Software install-any problems?

Installed tools; one issue-scoring in tools generate PNGs for Det curves-What version does your system support-4.2.4. Only one thing didn't pass main check-**Nancy to send a log**.

It's in your best interest to get it worked out. It says I should install 4.4 but couldn't find it for Suess Linux-11.1-4.4 is fairly recent-4.6 is the latest one.

Dimitra-We installed it yesterday, but I haven't test yet.

Also **send SSH keys as an attachment**

Chanel and Dorsi- we have some of the same problems but we haven't checked yet; will send email with problems

First time F4DE install takes work but reinstalls/updates are easier.

STD 06 DATA

NIST did not release the STM files for STT scoring but there are force aligned RTTM files. NIST will work on this.

Remember: There are two kinds of data distributions: (1) the audio and original transcripts, (2) the IndusDBs that contain system/scoring input files.

KW Search Time Documentation

Do we need to document search time for each KW?

 Search time is an optional requirement- nice if you can do it.

Voice Activity Detection and Scoring Group

For Voice activity detection – if we want to stay in BASE LR condition you have to use 80 hours of conversational training .

IndusDB README

IndusDB pack has a big readme describing the contents.

Future Open KW Search Telecons

April first 10 AM,

April 15th 10 AM and

maybe April 29th 10 AM