Appendix J: Build Pack Training Resources

This appendix describes the file and directory structure of the Surprise Language Build Pack materials to support defined training conditions and the tuning set. The data is released in the "Language Pack Definitions" (LPDEFs) located on the OpenKWS data website. The content below documents textual data for the build pack. The build pack audio will be distributed separately conforming to the Babel Data Specification (BDS) document posted at (http://www.nist.gov/itl/iad/mig/openkws15.cfm).

Speaker Demographics

Speaker demographics are provided for every audio file within the build pack. All of the information within the demographics file is usable for system development for all evaluation conditions.

- File: *./conversational/reference_materials/demographics.tsv*
- Format: Specified in the BDS

Phonetic Lexicon

The phonetic lexicon will be distributed as a resource to be used for contrastive conditions. **It will be released after all primary systems submissions are complete. See the evaluation schedule for the release (http://www.nist.gov/itl/iad/mig/openkws15.cfm) date and the Full Language Pack description below.**

Tuning set

A 3-hour selection of conversational data from a 10 hour build pack pool, BPack-Sub3, is provided for system parameter tuning. The tuning data is selected from a 10-hour subset of the build pack using segments (contiguous speech separated by 0.5 seconds of silence) as units to select. The tuning set is to be used only for meta-parameter tuning, e.g., acoustic/language model weights but not acoustic models themselves, for the VLLP and ALP training conditions. The following items are provided to define the tuning set:

- Audio Selection Definition: The selected audio is defined via an ECF file (Appendix A). Note: The selection is at the segment level and segments are not contiguous.
  - File: *./conversational/tuning/tuning.ecf.xml*
- Transcriptions: The transcripts are provided in two forms: Appen-style transcripts and Scoring STM transcripts. The Appen-style transcripts are the authoritative transcripts. The transcript for each audio file is in a separate file. The transcripts deviate from the Appen 'norm' in that both segment begin and end times are provided so that the segment times are fully specified. Segments for which transcription is not provided use the /(<untranscribed>)/ tag. The Scoring STM are provided for convenience to evaluate STT system components.
  - Appen-Style Transcript Directory: *./conversational/tuning/tuning transcripts*
  - Scoring STM File: *./conversational/tuning/tuning.stm*

Very Limited Language Pack

Transcripts for the VLLP condition consists of 3-hours selected from the build pack without any scripted materials. The data is selected from the 30-hour subset of the build pack pool, BPack-Sub2, using segments (contiguous speech separated by 0.2 seconds of silence) as units to select. The following items are provided to define the VLLP data:

- Audio Selection Definition: The selected audio is defined via an ECF file (Appendix A). Note: The selection is at the segment level and segments are not contiguous.
  - File: *./conversational/VLLP/VLLP.training.ecf.xml*
- Untranscribed Audio Definition: The training material available for systems training is defined via an ECF file. The usable build pack data excludes the Tuning set
  - File: *./conversational/VLLP/VLLP.untranscribed.ecf.xml*
- Transcriptions: The transcripts are provided in two forms: Appen-style transcripts and Scoring STM transcripts. See the description of the items in the Tuning transcripts above.
  - Appen-Style Transcript Directory: *./conversational/VLLP/VLLP transcripts*
  - Scoring STM File: *./conversational/VLLP/VLLP.stm*

Active Learning Limited Language Pack

The ALP is a 2-phase system build evaluation condition for Babel performers only. In phase 1, NIST provides participants with 1 hour of seed transcript excerpts (1/3 of the VLLP transcripts) that are used to identify, via automatic methods, an additional 2 hours of training audio for which the team requests additional training transcripts. In Phase 2, the participant's KWS system is built using the seed transcripts and the NIST-provided additional 2 hours of training transcripts that were requested. No scripted material will be provided with the ALP.

In Phase I, the following items are provided by NIST:

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: .

Jonathan Fiscus 2/3/2015 9:29 AM
Moved down [1]: <#> The lexicon is provided in the original Appen form per the BDS. The following items are provided by NIST:

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: 2

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: /[

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted:]/

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: in

Jonathan Fiscus 2/3/2015 9:30 AM
Deleted: KWS15-evalplan-v04.docx

- Audio Selection Definition: The selected audio is defined via an ECF file (Appendix A). Note: The selection is at the segment level and segments are not contiguous.
  - File: *./conversational/ALP/ALP.phase1.training.ecf.xml*
- Untranscribed Audio Definition: The training material available for systems training is defined via an ECF file. The usable build pack data excludes the Tuning set.
  - File: *./conversational/ALP/ALP.phase1.untranscribed.ecf.xml*
- Transcriptions: The transcripts are provided in two forms: Appen-style transcripts and Scoring STM transcripts. See the description of the items in the Tuning transcripts above.
  - Appen-Style Transcript Directory: *./conversational/ALP/ALP.phase1.transcripts*
  - Scoring STM File: *./conversational/ALP/ALP.phase1.stm*
- Active selection audio pool: The system will use the above resources to request transcripts for 2 additional hours from the audio defined via an ECF file.
  - File: *./conversational/ALP/ALP.phase1.selectionPool.ecf.xml*

To prepare for Phase 2, each team participating in the ALP condition will send NIST their requested segments. The requested segments are ordered by priority in a comma separated value (CSV) file containing three columns and the header:

filename,beginTime,endTime.

NIST will use force aligned, human transcripts to 'build' the team's Phase 2 transcripts by expanding the system-provided segmentations to the nearest silence gap of greater than 0.2 seconds. The following items will be provided by NIST to finish the ALP resources:

- Phase 2 Transcriptions: The transcripts are provided in two forms: Appen-style transcripts and Scoring STM transcripts.
  - Appen-Style Transcript Directory: *./conversational/ALP/ALP.phase2.transcripts*
  - Scoring STM File: *./conversational/ALP/ALP.phase2.stm*

Full Language Pack

The FullLP pack includes all available transcripts for the build pack. The amount of available transcripts changes by year: 80, 60, and 40 hours for 2013, 2014, and 2015+. The following items are provided by NIST:

- Transcriptions: The transcripts are provided as original Appen transcripts.
  - Appen Transcript Directory: *./conversational/training/transcription*
  - Appen Transcript Directory: *./scripted/training/transcription*
- Scripted demographics: *./scripted/reference_materials/demographics.tsv*
- Phonetic Lexicons: The lexicon is provided in the original Appen form per the BDS. The following items are provided by NIST:
  - File: *./conversational/reference_materials/lexicon.txt*
  - File: *./scripted/reference_materials/lexicon.txt*

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: in

Jonathan Fiscus 2/3/2015 9:29 AM
Deleted: *untranscribed*

Jonathan Fiscus 2/3/2015 9:29 AM
Formatted: List Paragraph, Indent: Left: 0.19", Hanging: 0.19", Bulleted + Level: 1 + Aligned at: 0.75" + Indent at: 1"

Jonathan Fiscus 2/3/2015 9:29 AM
Moved (insertion) [1]

Jonathan Fiscus 2/3/2015 9:29 AM
Formatted: Font:Times New Roman, 9 pt

Jonathan Fiscus 2/3/2015 9:29 AM
Formatted: Indent: Left: 0.38", Hanging: 0.19"

Jonathan Fiscus 2/3/2015 9:30 AM
Deleted: KWS15-evalplan-v04.docx