Open Eval KW Search Telecon 04 01 2013

	jonathan.fiscus@nist.gov
Attendees:	"Michel, Martial" <martial.michel@nist.gov>; "Mary P Harper" <mary.p.harper@ugov.gov>; "Ronnie F. Silber" <ronnie.f.silber@ugov.gov>; CHEN; Soon on; Singapore; Adam Janin; Rich Schwartz BBN; Andrew Rosenberg; Lydia, IBM; Stavros, San Gee, JHU; Vergyri, Dimitra

Agenda Items:

Update on scoring process - ECF Times derived from transcripts not waveforms. Transcripts are sometimes incorrectly longer than audio.

- Editorial Updated: Fixed in the 20130402 IndusDB.

Issue: The 107 STM file isn't sorted. If you use sclite, the file needs to be sorted- will make a new release of the Indus DB.

- Editorial Updated: Fixed in the 20130402 IndusDB.

Note: The STT scorer accepts CTM files rather than RTTM. Will fix Eval Plan to make this clear.

Note: There are some problems with validating CTM for STT- The next version of F4DE will have a fix.

Questions:

What is the frequency range of the data: All audio (including scripted) is 8Khz telephony data. So the range os 0-4Khz. The band limits are unknown and variable.

How many keywords should we expect for 75 hrs of eval data? Many more than 200.

Are people interested in sharing KWs for training? No.

Will OOVs will not be in lexicon? Are foreign KWs possible? We discussed the ABH process. If they show up in transcripts then they're part of language and possible keywords.

Other info:

- Can't use other resources in base LR like English corpora.
- Lexicon: White space used to mark words boundaries. Canonical Vietnamese tends to be single syllables. Imported words are poly syllabic - written in ASCII
- Case is removed for scoring

Future Open KW Search Telecons : April 15th 10 AM and April 29th 10 AM

-----Teleconference Access Numbers:-----
Toll Free: 877-710-9621
Toll: 1-203-566-5621

Passcode: 1704299

Schedule: Participants will receive Eval Pack Apr. 22. Evaluation Pack will be sent in advance, and then Password will be released.