The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan

(Version 4, Updated 02/25/2003)

1 INTRODUCTION

The goal of this document is to define the evaluation tasks, performance measures, and test corpora to support the 2003 Rich Transcription Spring (RT-03S) evaluation. Rich Transcription (RT) is broadly defined to be a fusion of speech-to-text (STT)[1] technology and metadata extraction (MDE) technologies which will provide the basis for the generation of more usable transcriptions of human-human speech for both humans and machines. This is the second evaluation in a series which seeks to identify, define, and evaluate individual RT component tasks as well as certain useful integrations. This series provides the evaluation mechanism for some of the research challenges in the DARPA Effective, Affordable, Reusable Speech-to-text (EARS) Program.[2] Note, however, that in addition to EARS contractors, this evaluation is open to all interested volunteers. All participants will be permitted to attend the RT-03 Spring Workshop which will follow the evaluation.

This evaluation supports two major tasks:

> **Speech-to-Text Transcription (STT)** – This evaluation targets the conversion of recordings of speech to strings of lexical tokens.
>
> **Metadata Extraction (MDE) – "Who Spoke When" Diarization** – This evaluation targets the identification and classification of speakers within recordings of speech.

In future evaluations, the STT task will remain constant but additional MDE tasks will be defined. This fall, for instance, we expect to add evaluations of structural units and disfluencies.[3] This document describes only the tasks for the RT evaluation this spring.

2 BACKGROUND

Beginning in the early 1980's, evaluation of automatic speech recognition (ASR) stabilized on the current performance measure of word error rate (WER). This measure scores ASR performance using a caseless lexicalized form of ASR output known as the "standard normalized orthographic representation" (SNOR) format.[4] The WER is defined as the sum of all ASR output token errors divided by the number of scoreable tokens in a reference transcription of the test data. There are three types of errors, these being namely tokens that are missed (deletion errors), inserted (insertion errors), and incorrectly recognized (substitution errors).[5]

While the traditional STT evaluations have helped to provide a mechanism for evaluating word accuracy, it is clear that words alone are insufficient in formulating a transcription of speech which is readable by humans and understandable by machines. A verbatim transcription of the speech stream into a string of lexical tokens yields a transcript that is often extremely difficult to understand. This is because spoken language is much more than just a string of lexical tokens. It contains information about the speaker, prosodic cues to the speaker's intent, and much more. Spoken language also contains disfluencies, which speakers correct and which textual renderings should delete. All of this makes the task of rendering spoken language into text a great challenge, especially with less-than-perfect ASR performance.

Solving these problems is the challenge that the EARS program takes as its objective and what the RT evaluation series seeks to assess – namely to develop technology that transforms spoken language into a form that is maximally informative. This requires new approaches to acoustical modeling and insightful models of disfluencies, dialogue and other relevant speaker behaviors.

3 EVALUATION TASKS

Separate evaluations are defined for each of the RT tasks. These tasks, and the evaluations of them, are defined to be as independent of each other as possible.

A major change in the traditional NIST evaluation task definitions this year is the addition of the requirement to transcribe periods of overlapping speech. This is accommodated by an output structure definition that supports multiple transcription streams for different speakers as well as different channels.

3.1 STT

The STT task is similar to previous ASR "Hub-4" and "Hub-5" evaluations, with some new additions which support the classification of output tokens, confidence measures, and

[1] formerly known as automatic speech recognition (ASR)

[2] The EARS research effort is dedicated to developing powerful new speech transcription technology that provides substantially richer and more accurate transcripts than are currently possible. The research focus is on natural, unconstrained speech from broadcasts and telephone conversations in a number of languages. The program objective is to create core enabling technology suitable for a wide range of advanced applications.

[3] It is currently planned that in addition to the set of new MDE tests, the Fall RT tests will include STT on meeting room data only and will not include STT on CTS or BNews data. (However, some EARS sites may be required to run a new set of Progress tests in the Fall if deemed necessary by the sponsor.)

[4] Since some languages' written forms are not word-based, this concept has been extended to cover lexemes – a representation of a written unit of meaning within a language. Thus, this document frequently refers to lexemes, lexical tokens, or tokens rather than words. For English, these terms may be treated more or less equivalently.

[5] Underlying the tabulation of errors is a requirement to align the tokens in the system output transcript with the tokens in the reference transcript. Traditionally, this has been done using dynamic programming so that the WER is minimized.

(optionally) speaker assignment. The required and optional aspects of the task are detailed in Section 4.1.

3.2 MDE

The metadata extraction task is a new task and one that is at an early stage of conceptual development. With the exception of the Speaker Diarization Who Spoke When task (see below), the RT-03 MDE task evaluations have been deferred until the Fall RT-03F evaluation. A second evaluation plan will be developed at a later date to describe those tasks.

3.2.1 Diarization – "Who Spoke When"

Diarization is the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources. These sources can include particular speakers, music, background noise sources, and other signal source/channel characteristics. For RT-03S, the diarization task will be limited to just the speaker segmentation "who spoke when" task including speaker gender classification. Note that upcoming diarization tasks, such as the "who said what" task planned for inclusion in the fall evaluation will focus on annotating the token stream with speaker information.

For the "who spoke when" task, small pauses in a speaker's speech, specifically of less than 0.3 seconds, are not considered to be segmentation breaks. Pauses of less than 0.3 seconds should be bridged into a single continuous segment. Although somewhat arbitrary, the cutoff value of 0.3 seconds has been determined to be a good approximation of the minimum duration for a pause in speech resulting in an utterance boundary. Systems should consider vocal noise (laugh, cough, sneeze, breath, lipsmack) to be silence in constructing segment boundaries.[6] See the documents, *Reference Data Cookbook for the Who Spoke When Diarization Task* and *Guidelines for RT-03 Transcription,* published on the EARS website (http://ears.ll.mit.edu/) for specific information about how the diarization reference is created and how specific data types (such as speaker-attributable noise) should be processed.

Required speaker attributes to be recognized include only the speaker type, namely one of "adult_male", "adult_female", "child", or "unknown". These labels must be consistently applied to all segments attributed to a particular speaker.

Note that systems may make use of the output of a word/token recognizer (or any form of automatic signal processing) in performing this task. The approach used should be clearly documented in the task system description.

4 Performance Measures

Separate performance measures are defined for each of the major EARS tasks.

4.1 STT

The STT performance measure is essentially the same as the traditional NIST ASR WER measure using the NIST SCLITE software. However, this year an effort will be made to evaluate over all speech – including areas of overlap where two or more talkers are speaking simultaneously on the same channel. While the primary metric for the EARS Program will remain WER for non-overlapping speech, the new additional metrics including overlapping speech will provide a look to the future for domains which might present a significant overlap problem. To implement this, a means of aligning multi-stream transcripts will be developed. The general approach will be to map each speech segment in the STT output onto the reference data so as to yield the best overall WER.

Token string formatting:

- A single standardized spelling is required for scoreable lexemes, and the STT system must output this spelling in order to be scored as correct.[7] Homophones must be spelled correctly according to the given context in order to be considered correct. All tokens are to be generated according to Standard Normal Orthographic Representation (SNOR) rules:
- Whitespace-separated lexical tokens (for languages that use whitespace-defined words)
- Case insensitive alphabetic text (usually in all upper case)
- Spelled letters are represented with the letter followed by a period (e.g., "a. b. c.")
- No non-alphabetic characters (except apostrophes for contractions and possessives and hyphens for hyphenated words and fragments)

Note that in scoring, hyphenated words will be divided into their constituent parts. Thus, for scoring, a hyphen within a token will be treated as a token separator. A hyphen at either end of a token string indicating the missing part of a spoken fragment will be discarded.

System output generation:

The system output is token-based and is to include the following information for each recognized token: the name of the source file and channel processed, the beginning time of the recognized token, the duration of the recognized token, the string representation of the recognized token, a confidence probability, a token type, and a speaker identifier. The speaker information is optional, but is included to support STT/MDE fusion experiments. If no speaker information is generated, a value of "unknown" should be used for lexical token types and "null" for non-lexical token types. See Section 7.2.2 for specific formatting requirements. The following describes each possible system output token type:

- **lex** - a lexical token. All other token types listed below will be stripped from the system output prior to scoring. Therefore only tokens tagged as type lex in the system output will be aligned and scored.
- **frag** - a lexical fragment (optional). Note: A (optional) hyphen may also be used in the token string to indicate the

[6] However, special scoring rules will apply to areas containing vocal noise. See Section 4.2.1.

[7] Token spelling is determined by NIST by first consulting an authoritative reference – e.g., the American Heritage Dictionary (AHD) for English. Lacking an authoritative reference, the www is searched to find the most common representation. If no single form is dominant, then two or more forms will be permitted via an orthographic map file. As in previous years, a transcription filter and orthographic map file will be used on both the reference and hypothesis transcripts to apply rules for mapping common alternate representations to a single scoreable form.

missing (unspoken) part of the token, but the frag TYPE must also be used.

- **fp** - a filled pause (optional).
- **un-lex** - an uncertain lexical token normally used only in the reference (optional).
- **for-lex** - a "foreign" lexical token (optional) normally used only in the reference.
- **non-lex** - a non-lexical acoustic phenomenon (breath-noise, door-bang, etc.) (optional).
- **misc** - other annotations not covered in above (optional).[8]

Although systems aren't penalized (or rewarded) for outputting the optional types above, we encourage their output to support metadata experiments.

Reference token processing:

A Segment Time Marked (STM) scoring reference is generated from the human reference transcripts. Non-scoreable regions (such as untranscribed areas) are explicitly tagged in the STM file for exclusion from scoring. Note that this will contain contraction expansions if they have been annotated in the human reference. The tokens in the STM reference will be processed as follows:

- **lex** - Tokens of type lex are not specially tagged in the reference. As such, they are aligned and scored.
- **frag** - Tokens of type frag are tagged as both optionally deletable and fragments in the reference. Since tokens of this type will be stripped from the system output prior to scoring, they will simply contribute to the WER denominator. Note: In addition, if a system output token of type lex aligns with a frag in the reference, it is counted as correct if the reference frag token string is a substring of the system output token string.
- **fp, un-lex, for-lex** – Tokens of these types are tagged as optionally deletable in the reference. Since they will be stripped from the system output prior to scoring, they will simply contribute to the WER denominator.
- **non-lex** and **misc** – These token types are removed from the reference

GLM Processing:

Prior to scoring, both the reference and system output token strings will be transformed using a global map file (GLM). The GLM is intended to ensure that reference and hypothesis tokens which do not differ semantically are scored as correct. This is accomplished by transforming the token strings in both the reference and system output via a set of mapping rules. The GLM applies a set of rules to the system output which expands contractions to all possible expanded forms.

Note that GLM processing may result in the generation of several alternative token strings in the system output. It may also result in token strings being split into two or more strings. For example, contractions are mapped to their expanded form and compound words are split into their constituents. After GLM filtering, hyphens in both the system output and reference are transformed into token separators.

Scoring:

Once the pre-processing is complete, token alignment will be performed using a token-mediated alignment optimized for minimum word error rate. An overall STT error score will be computed as the average number of token recognition errors per reference token:

$$Error_{STT} = \left(\mathrm{N}_{\mathrm{Miss}} + \mathrm{N}_{\mathrm{FA}} + \mathrm{N}_{\mathrm{Err}}\right)/\mathrm{N}_{\mathrm{Ref}}$$

where

$\mathrm{N}_{\mathrm{Miss}}$ = the # of unmapped reference tokens,

N_{FA} = the # of unmapped STT output tokens,

$\mathrm{N}_{\mathrm{Err}}$ = the # of mapped STT output tokens with non-matching reference spelling per the token rules above, and

$\mathrm{N}_{\mathrm{Ref}}$= the maximum number of reference tokens[9]

As an additional optional performance measure, the confidence of a system in its transcription output will be evaluated. In order to do this, the system must attach a measure of confidence to each of its scoreable output tokens. This confidence measure represents the system's estimate of the probability that the output token is correct and must have a value between 0 and 1 inclusive. The performance of this confidence measure will be evaluated using the same normalized cross entropy score that NIST has been using in previous ASR evaluations.[10]

Conditioned Sub-Scoring:

STT WER performance statistics will be tabulated for the following conditions:

- **Language** – Performance will be measured separately for English, Chinese (Mandarin), and Arabic language data.
- **Source** – Performance will be measured separately for broadcast news sources and for telephone conversations.
- **CPU processing time** – See section 6.1 for processing time options and requirements.
- **Speaking conditions** – Performance will be measured separately for the following speaking conditions:
  - Non-overlapping speech. (primary metric for EARS)
  - Overlapping speech
  - All speech

4.2 MDE

4.2.1 Diarization – "Who spoke when"

The results for this task will be analyzed in a similar manner to the NIST 2002 speaker segmentation evaluation[11], albeit with a slightly different metric (see below).

[8] Any token which is to be excluded from scoring may be given this tag – including those for which specified types exist. However, where possible, sites are encouraged to use the supported types to enhance the usefulness of the data for MDE experiments.

[9] $\mathrm{N}_{\mathrm{Ref}}$ includes all scorable tokens (including optionally deletable tokens) and counts the maximum number of tokens where there are alternatives (as for possible contractions). Note that $\mathrm{N}_{\mathrm{Ref}}$ considers only the reference transcript and is not affected by tokens in the system output transcript, regardless of their type.

[10] http://www.nist.gov/speech/tests/rt/rt2003/doc/NCE.htm

[11] http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf

The reference segmentation for this task will be constructed per the *Reference Data Cookbook for the Who Spoke When Diarization Task* published on the EARS website (http://ears.ll.mit.edu/). Note that special procedures will be applied to the construction of segments containing vocal noise.

Note that a UEM[12]-formatted file will be used to eliminate certain regions from scoring[13], including untranscribed regions and areas surrounding vocal noise -- extending from the vocal noise to the closest segment or word boundary in both directions.

Speaker Segmentation Diarization Scoring:

In order to measure performance, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs will be computed. The measure of optimality will be the aggregation, over all reference speakers, of time that is jointly attributed to both the reference speaker and the (corresponding) system output speaker to which that reference speaker is mapped. This will always be computed over all speech, including regions of overlap. Mapping is subject to the following restrictions:

- Mapping will be one-to-one, meaning that each reference speaker will map to at most one system output speaker, and each system output speaker will map to at most one reference speaker.
- Mapping of speakers will be computed separately for each speech data file.

Unlike in previous segmentation evaluations, areas of overlapping speech will be evaluated. However, the primary metric will be based on non-overlapping speech only. In addition, since segment times are assumed to be correct in the reference in this evaluation, no time collars will be employed to forgive timing errors in the reference.

Speaker detection performance will be expressed in terms of the miss and false alarm rates that result from the mapping. The accuracy of recognition of speaker attributes will be computed for successfully detected speakers (i.e., for mapped speakers) and separately for all system output speakers.

An overall time-based speaker diarization error score will be computed as the fraction of speaker time that is not attributed correctly to a speaker. This will be the **primary metric** for speaker segmentation diarization:

$$Error_{SpkrSeg} = \frac{\sum_{\substack{\text{all}\\ \text{segs}}} \left\{ \mathrm{dur}(seg) \cdot \left(\max\left(\mathrm{N}_{\mathrm{Ref}}(seg), \mathrm{N}_{\mathrm{Sys}}(seg) \right) - \mathrm{N}_{\mathrm{Correct}}(seg) \right) \right\}}{\sum_{\substack{\text{all}\\ \text{segs}}} \left\{ \mathrm{dur}(seg) \cdot \mathrm{N}_{\mathrm{Ref}}(seg) \right\}}$$

where the speech data file is divided into contiguous segments at all speaker change points and where, for each segment, *seg*:

$$\begin{aligned} \mathrm{dur}(seg) &= \text{the duration of } seg, \\ \mathrm{N}_{\mathrm{Ref}}(seg) &= \text{the \# of reference speakers speaking in } seg, \\ \mathrm{N}_{\mathrm{Sys}}(seg) &= \text{the \# of system speakers speaking in } seg, \\ \mathrm{N}_{\mathrm{Correct}}(seg) &= \text{the \# of reference speakers speaking in } seg \\ &\quad \text{for whom their matching (mapped) system} \\ &\quad \text{speakers are also speaking in } seg. \end{aligned}$$

The numerator of the overall diarization error score represents speaker diarization error time, and it can be decomposed into speaker time that is attributed to the wrong speaker, missed speaker time, and false alarm speaker time.

Speaker time that is attributed to the wrong speaker (called speaker error time) is the sum of the following over all segments:

$$\mathrm{dur}(seg)^* \ \{\min(\mathrm{N}_{\mathrm{Ref}}(seg), \mathrm{N}_{\mathrm{Sys}}(seg)) - \mathrm{N}_{\mathrm{Correct}}(seg)\}.$$

Missed speaker time is the sum of the following over only segments where more reference speakers than system speakers are speaking:

$$\mathrm{dur}(seg)^*(\mathrm{N}_{\mathrm{Ref}}(seg) - \mathrm{N}_{\mathrm{Sys}}(seg)).$$

False alarm speaker time is the sum of the following over only segments where more system speakers than reference speakers are speaking:

$$\mathrm{dur}(seg)^*(\mathrm{N}_{\mathrm{Sys}}(seg) - \mathrm{N}_{\mathrm{Ref}}(seg)).$$

Word-based counterparts to the time-based speaker diarization error score, and to each of its three parts, are calculated by using word counts instead of time, counting the number of reference words whose midpoint time falls in the segment, (midpoint time is the start time of the word plus half its duration).

In areas of overlap (that is, segments where more than one reference speaker is speaking), note that the duration of the segment is attributed to all the reference speakers who are speaking in the segment thus counting the time more than once. But since the reference data tells us which speaker actually spoke each reference word, we attribute each word to its actual speaker, and in areas of overlap this means the words are not counted more than once

Speaker Type (Gender) Diarization Scoring:

The diarization "who spoke when" scoring program can be run in a mode that uses the speaker type (adult_male, adult_female, child, or unknown) as the speaker ID. In this mode, the program will bypass the algorithm to compute an optimum mapping of reference speakers to system output speakers, as the set of possible speaker types is known *a-priori* and no mapping is required. As a result, more of the time and words are likely to be mapped. The output in this mode will include the same time-based and word-based metrics described above, but will also include confusion matrices for the speaker types.[14] The **primary metric** for

[12] UEM *(Unpartitioned Evaluation Map*) and is a file format used to create an index specifying time regions within a recorded waveform.

[13] This UEM-formatted file should not be confused with the UEM-formatted *test index* file used to specify the test material. The material specified in this UEM *score* file will be a subset of the test material specified in the UEM test index file if reference transcriptions don't exist for some of the material (such as for commercial segments in broadcast news).

[14] These speaker type confusion matrices are always generated by the program, both for speaker segmentation scoring and speaker

speaker type diarization is the same as indicated above for speaker segmentation diarization.

As an additional optional performance measure, the confidence of a system in its diarization output may be evaluated. To support this, a system would have to attach a measure of confidence to each of its output speaker segments. While this may be required for future evaluations, it is optional for the RT-03 Spring evaluation. This confidence measure represents the system's estimate of the probability that the speaker of this segment is correctly assigned.[15] The performance of this confidence measure will be evaluated using essentially the same normalized cross entropy scoring procedure as used to evaluate the token confidence measure for STT.

Conditioned Sub-Scoring:

MDE Who Spoke When Diarization segmentation statistics will be tabulated for the following conditions:

- **Source** - Performance will be measured separately for broadcast news sources and for telephone conversations.
- **Type**
  - Speaker ID
  - Speaker Gender
- **Speaking** conditions – Performance will be measured separately for the following speaking conditions:
  - Non-overlapping speech. (primary metric for EARS)
  - Overlapping speech
  - All speech

5 CORPUS SUPPORT

5.1 TRAINING AND DEVELOPMENT TEST DATA

Corpora to support the training and development of the STT and MDE tasks specified in this document are provided as indicated in Table 1. These corpora are evolving over time and, as such, this information is likely to change. Note that **all** material used in **any** way for training and development for the broadcast news recognition tasks must predate the test epoch (February 2001) as specified in Section 7.1.2.

5.2 EVALUATION TEST DATA

The broadcast news and conversational telephone corpora for this evaluation have been collected, transcribed, and annotated by the Linguistic Data Consortium (LDC). This data is outlined in Table 1.

Although systems are required to run over large contiguous segments of speech, please note that certain regions of the broadcast news corpora are not transcribed or annotated (and will therefore not be scored). These untranscribed regions consist of commercials, reporter chit-chat outside of the context of a story, station identifications, public service announcements, promotions for upcoming broadcasts and long musical interludes. All remaining material which has been transcribed and annotated will be scored. Detailed documentation regarding the creation of this data may be found on the EARS website at http://ears.ll.mit.edu/.

5.2.1 SPEECH-TO-TEXT TRANSCRIPTION (STT)

English STT:

The STT English evaluation will be conducted on two different data sets: a "progress" data set to be used only by EARS contractors and affiliates and a "current" data set to support tests open to all participants. The Progress Set will remain fixed and reused over time. A fresh Current set will be created for each evaluation cycle:

- The *Current* English data set is intended to represent the real world problems of telephone signal quality and language evolution. The broadcast news test set will comprise approximately 3 hours of broadcast news. The conversational telephone test set will consist of two distinct 3-hour conversation subsets. These subsets will be drawn from the existing SWBD-Cellular collection and new "Fisher" collection, respectively. (These subsets are to be processed as a single set and the origin of the data is to be unknown to the systems.) This duality is intended to permit comparison of performance for the new Fisher data to that for known corpora. Thus the total amount of conversational speech will be 6 hours. The conversational telephone speech data for later evaluations will consist of Fisher data only. *These tests (and data) are open to all participants.*
- The *Progress* data set is an English language data set. This data set will be reserved for measuring EARS year-to-year progress and will not be available for any purpose other than testing. It will comprise approximately 3 hours of broadcast news and 3 hours of Fisher Corpus telephone conversations. Please refer to the specific rules governing the use of this data in a document to be published on the EARS website at http://ears.ll.mit.edu/. *These tests (and data) are only open to EARS contractors and affiliates.*

The evaluation test data will consist of nominally 30-minute uninterrupted excerpts of news broadcasts, taken from the beginning of the broadcast and nominally 5 minute telephone conversation excerpts taken from the telephone conversations. The news broadcasts will be endpointed to the nearest story boundary and the telephone conversations will be endpointed to the nearest turn boundary. Therefore, each broadcast news excerpt may be slightly smaller or larger than 30 minutes and each telephone conversation may be slightly smaller or larger than 5 minutes. Systems will be expected to process the specified excerpts in their entirety, even though they may contain some material (such as commercials and untranscribable passages) which will be excluded in scoring.

Non-English STT:

A set of tests similar to the English Current tests will be conducted using Chinese (Mandarin) and Arabic broadcast news and conversational telephone speech. Each test set is 60 minutes in length. The Chinese test sets consists of 12-minute continuous excerpts from 5 broadcast news sources and 12 5-minute telephone conversations. The Arabic test sets consist of 30-minute continuous excerpts from 2 broadcast news sources and 12 5-minute telephone conversations. The broadcasts for both languages are from February 2001 (the same as the English broadcast news test

type scoring. However, they will differ for segmentation and type scoring since they are based on different mappings.

[15] The confidence measure represents the confidence in speaker assignment only. It should exclude consideration of the correctness of other attributes such as speaker type and segment times.

epoch). No progress tests are being implemented for the non-English tests. *These tests (and data) are open to all participants.*

5.2.2 Metadata Extraction (MDE)

The evaluation of the Speaker Diarization task will be conducted on a portion of the English "Current" BNews and CTS STT data sets. The MDE BNews test set will be composed of the first half, temporally, of the news broadcasts used for the STT tests. Likewise, the MDE CTS test set will be composed of 90 minutes of half of the CTS data used in the STT tests. For the CTS data, it will be divided so as to provide an even distribution of the data demographics both across the test set and within the remaining unused material. Note, however, that the CTS data will not be divided within a conversation. This split is being implemented so that the remaining, unused data may be used for the RT-03 Fall MDE evaluations.

6 Evaluation Conditions

There are many different conditions under which system performance may be evaluated. This section identifies those conditions for which performance will be computed and, of those, which are to be designated as the "primary" evaluation conditions.

As a general rule, for both STT and MDE evaluation, adaptive use of all of the evaluation data (and legal training data) will be allowed. For adaptive systems, the order of presentation and use of the data can affect results and is important. Therefore, for the broadcast news data, the evaluation data will be presented in chronological order and systems must process the data in this order. During the processing of each data file, the data in that file may be used to adapt and otherwise modify the processing system. Subsequent data files, however, may not be accessed before processing and output is complete for the current file. Time sequence is of little consequence with regard to the conversational telephone data, so no such constraints apply for that data.

6.1 STT

The following evaluation conditions will be supported:

- **Language:**
  - STT may be implemented on the following languages: English, Chinese (Mandarin), and Arabic
- **Domain**:
  - Broadcast news and conversational telephone speech
- **Input**:
  - Only one input condition is supported for each of the STT tasks – namely the speech (audio file only) input condition.
- **Processing time**:
  - Performance will be measured separately for three different CPU processing time factors[16], namely ≤1X, ≤10X, and unlimited. For the EARS 2003 evaluation, the primary evaluation condition for broadcast news will be ≤10X and the primary evaluation condition for telephone conversations will be unlimited CPU processing time.
- **System version**:
  - State-of-the-art system run
  - EARS sites will be running a mothballed version of their RT-02 systems on the Progress set. *These mothballed system tests are required for EARS contractors and affiliates only.*

6.2 MDE

6.2.1 Diarization – Who Spoke When

Evaluation will be performed separately for each file and for each channel within a file, and so unique determination of speakers will be limited to within-channel data. Therefore, tracking of speakers across channels and files is not required.[17]

The following evaluation conditions will be supported:

- **Language**:
  - English only
- **Domain**:
  - Broadcast news and conversational telephone speech
- **Input**:
  - MDE input will be the speech input. Any desired fully-automatic signal processing approaches may be employed (including the use of the output of an STT system[18]).
  - MDE input will include, along with the speech input data, the reference transcriptions (including the time marks). Only the reference token and token-time information may be used. No reference speaker information may be used for diarization tasks. This condition is an optional contrast. Condition 1 must also be implemented if this condition is implemented.

7 Participation Instructions

Participation is encouraged for all those who are interested in one or more of the RT-03 tasks. All participants must, however, agree to completely process all of the data for at least one task. For STT this means that, as a minimum, either all of the broadcast news data or all of the conversational telephone data for one language must be processed. *(however, EARS contractors*

[16] For purposes of this evaluation, the time to be reported is the actual "wall clock" time it takes to process all channels of the recorded speech (including I/O) on a single CPU. Therefore the real-time factor equals the above processing time divided by the duration of the signal in the processed recordings (across all channels for multi-channel recordings). For example, a 1-hour news broadcast processed in 10 hours would be 10X realtime (regardless of whether the broadcast is stereo or monaural). And a 5-minute telephone conversation processed in 50 minutes would also be 10X realtime (regardless of whether the signal is a 4-wire/2-channel signal or a 2-wire/1-channel signal).

[17] Neither is it prohibited, however. There are no restrictions against the use of cross-channel, cross-file or even cross-corpus information in the speaker diarization process. This includes both the evaluation data and all legal training data. Note however that the processing of evaluation data is subject to the causality constraint described at the beginning of section 6.

[18] Knowledge of the lexemes that are being spoken is an important part of performing MDE, and so speech recognition is part of MDE. The ASR engine that supplies this information may of course be a separate module, provided either by an in-house system or from other sources. Cross-site collaboration is encouraged.

much process both). For MDE - diarization, this means that as a minimum, the speech-input-only processing condition must be implemented.

As a condition of participation, all sites must agree to make their submissions (system output, system description, and ancillary files) available for experimental use by other research sites. (For example, NIST will make the Current English STT output and results available for use in MDE experiments.) Further, submission of system output to NIST constitutes permission on the part of the site for NIST to publish scores and analyses for that data including explicit identification of the submitting site and system.

7.1 Processing Rules

7.1.1 Rules that apply to all evaluations

All processing for all tasks must be fully automatic. No manual intervention of any kind is permitted. Systems will be provided with recorded waveform files and an index file specifying the speech files and regions within them to be processed. Conversational telephone speech test data will be provided in 2-channel files, and both channels must be processed. Broadcast news speech test data will be presented in single channel files. Each conversation and each news broadcast excerpt to be processed will be presented in a separate file. While entire broadcast and conversation files will be distributed, only the material specified in the UEM test index file for the experiment to be run is to be processed. Material outside of the times specified in the UEM test index file is not to be used in any way (e.g., for adaptation).

7.1.2 Additional rules for processing Broadcast News

News-oriented material (audio, textual, etc.) generated during or after the test epoch beginning February 01, 2001 **may not be used in any way for system development or training.** Broadcast news material must be processed in the chronological order of the date/time of the original broadcast. Although automatic adaptation may be performed using previously-processed material, systems may not “look ahead” in time at later recordings. As such, processing must be complete on a particular broadcast news test file before moving on to the next file. However, any form of within-file adaptation is permitted and systems may look backwards in time at previously-processed files. The show identity and original broadcast date are allowable side information that systems may use. Therefore, systems may make use of show-dependent models.

7.1.3 Additional rules for processing conversational telephone speech

Conversational telephone speech may be processed in any order and any form of automatic within-conversation and cross-conversation adaptation may be employed. No side information is provided for telephone conversations. Also, unlike last year, no manual or automatic segmentation will be provided, although systems may make use of segmentation outputs donated from other sites.

7.1.4 Additional rules for performing the STT task

The same system must be used to process both the Progress and Current Test sets. **Please note that to ensure the integrity of the Progress Test Set, special rules governing the use (and disposal) of this data must be strictly observed. These are specified in a document to be published at the EARS evaluation website at http://ears.ll.mit.edu/.**

Note that **all** of the constraints specified for the English STT tests regarding training, adaptation, and processing also apply to the Non-English STT tests

7.2 Data Formats

7.2.1 Test data

For practicality, the recorded waveform files to be processed will be distributed on CD-ROM and the corresponding indices, annotations, and transcripts will be made available via the Web or FTP using an identical directory structure. In addition, the Progress Test Set will be distributed separately from the other test material. Although different data sets will be distributed separately, the following directory structure will be used to both distribute the test data and accumulate and re-distribute the system output from the tests:

Directory	Description
indices/	index files containing the list of files and times to be processed for particular experiments
audio/	audio files
input/<EXP-ID>/	ancillary data including reference annotations for various experiments – must be used in accordance with instructions for that experiment
output/<EXP-ID>/	system output submissions – will be made available as received for integration tests [19]
reference/	reference transcripts, annotations, and MS-wav files for post-evaluation scoring and analyses

Note: EXP-ID specifies a unique identifier for each experiment and is defined in 7.3.1.

For clarity, the “audio/” and “reference/” directories are subdivided into <DATA>/<LANG>/<TYPE> subdirectories:

[dev|eval03|prog]/[english|mandarin|arabic]/
[bnews|cts]/

The “indices/” directory contains a set of UEM test index files specifying the waveform data to be evaluated for each EXP-ID condition supported in this evaluation as described in 7.3.1 and are named <EXP-ID>.uem with the special site code “expt”. A separate .uem file will be provided for each experiment for each supported <DATA>, <LANG>, and <TYPE>. Only the waveform data specified in these files should be processed for the given experimental condition. Corresponding ancillary data for some control conditions is given in the “input/” directory under subdirectories with the same EXP-ID. These files contain new-line-separated records and whitespace-separated fields of the form:

[19] However, no data regarding the Progress tests will be posted.

<FILE><SP><CHANNEL><SP><BEGIN-TIME><SP>
<END-TIME><NEW-LINE>

where,

<SP> is whitespace

<FILE> specifies the path and filename of the waveform file to be processed

<CHANNEL> specifies the channel within the waveform file to be processed

<BEGIN-TIME> and <END-TIME> specify the time region within the specified file to be processed.

For example:

The index file expt_03_stt10x_dev_eng_cts_spch_1.uem will contain:

```
.

.

audio/dev/english/cts/sw_47620.sph 0 0 291.34

.

.
```

7.2.2 STT OUTPUT FORMAT

The EARS 2003 STT output format will be the CTM format (.ctm filetype) specified in previous ASR evaluations with a modification to support the explicit typing of output tokens (scored as indicated in Section 4.1) including non-lexical information (not scored).[20] Each output file is to begin with two special comment lines specifying the experiment run and inputs used. These lines must appear at the beginning of the file and are to be formatted as follows:

The first line is a special comment specifying the experiment as defined in Section 7.3.1 (EXP-ID) and is of the form:

;; EXP-ID: <EXP-ID> [21]

For example,

;;EXP-ID: bbn_03_stt10x_eval03_eng_cts_spch_1

The second line is a special comment specifying the inputs used (excluding speech files) and is of the form:

;;INPUTS: <FILE1>,<FILE2>, ... <FILEn>

For example,

;;INPUTS: sw_47620.mdtm

Unless path information is included, these files are assumed to be under the same directory as the CTM file.

Note that for purposes of scoring, all lines beginning with ";;" are considered comments and are ignored. Blank lines are also ignored.

The header comments are followed by a list of CTM records. For those who've used the CTM format in the past, the CTM record format to be used for the EARS 2003 evaluation has been modified to add two additional attributes for each token and to eliminate special characters in the token field. These new CTM attributes are a token type and a speaker identifier. See the list below for the specific supported token types.

The CTM file format is a concatenation of time mark records for each output token in each channel of a waveform. The records are separated with a newline. Each field in a record is delimited with whitespace. Therefore, field values may not include whitespace characters. Each record follows the following BNF format:

CTM-RECORD :== <SOURCE><SP><CHANNEL><SP>
<BEG-TIME><SP><DURATION><SP><TOKEN><SP>
<CONF><SP><TYPE><SP><SPEAKER><NEWLINE>

where

<SP> is whitespace.

<SOURCE> is the waveform basename (no pathnames or extensions should be included).

<CHANNEL> is the waveform channel: "1", "2", etc. This value will always be "1" for single-channel files.

<BEG-TIME> is the beginning time of the token. This time is a floating point number, expressed in seconds, measured from the start time of the file. [22]

<DURATION> is the duration of the token. This time is a floating point number, expressed in seconds. [22]

<TOKEN> is the orthographic representation of the recognized word/lexeme or acoustic phenomena. For English, this is represented as a string of ASCII characters. (a token in the context of a non-English test might be represented in Unicode or some other special character set.) Token strings are case insensitive and may contain only upper or lowercase alphabetic characters, hyphens (-), and apostrophes (') only. No special characters are to be included in this field to indicate the type of token. Rather, the "TYPE" field is to be used to indicate the token type. Note however that a hyphen may be used for fragments to indicate the missing/unspoken portion of the fragment. However, the "frag" TYPE must still be used.

<CONF> is the confidence score, a floating point number between 0 (no confidence) and 1 (certainty). A value of "NA" is to be used when no confidence is computed and in the reference data. [23]

[20] This information is being encouraged as potential input to MDE systems.

[21] The EXP-ID will be ignored for all non-evaluated files (including CTM files used for MDE experiments). You may use whatever value you wish for these files. However, all files to be scored must have properly formatted EXP-IDs.

[22] A required time accuracy for BEG-TIME and DURATION is not defined, but these times must provide sufficient resolution for the evaluation software to align tags with the proper token in the reference when time-alignment-based scoring is used. This alignment can be problematic in the case of quickly-articulated adjoining words. Therefore, systems should produce time tags with as much resolution as is reasonably possible. Note that the word with the shortest duration in the MDE development test set is 15 ms.

[23] STT systems are required to compute a confidence for each scoreable token output for this evaluation. The "NA" value may be used only for non-scoreable tokens.

<TYPE> is the token type. The legal values of <TYPE> are "lex", "frag", "fp", "un-lex", "for-lex", "non-lex", "misc", or "noscore". See Section 4.1 for details on generation and scoring rules for each of these types.

- lex is a lexical token.
- frag is a lexical fragment. Note: A (optional) hyphen may also be used in the token string to indicate the missing (unspoken) part of the token, but the frag TYPE must also be used.
- fp is a filled pause.
- un-lex is an uncertain lexical token normally used only in the reference.
- for-lex is a "foreign" lexical token normally used only in the reference.
- non-lex is a non-lexical acoustic phenomenon (breath-noise, door-bang, etc.)
- misc is other annotations not covered in above. [24]
- noscore is a special tag used only in reference files for scoring to indicate tokens which should not be aligned or scored.

<SPEAKER> is a string identifier for the speaker who uttered the token. This should be "null" for non-speech tokens and "unknown" when the speaker has not been determined.

Included below is an example of STT system output:

```
7654 1 11.34 0.2 YES 0.763 lex 1
7654 1 12.00 0.34 YOU 0.384 lex 1
7654 1 13.30 0.5 C- 0.806 frag 1
7654 1 17.50 0.2 AS 0.537 lex 1
:
7654 2 1.34 0.2 I 0.763 lex 2
7654 2 2.00 0.34 CAN 0.384 lex 2
7654 2 3.40 0.5 ADD 0.806 lex 2

7654 2 3.70 .2 door-bang 0 non-lex null
7654 2 7.00 0.2 AS 0.537 lex 2
:
```

7.2.3 MDE Output Format

The EARS 2003 MDE output format (.mdtm filetype) is a tabular text file with whitespace-separated fields and new-line-separated records. Lines beginning with a double semicolon are treated as comments.

Each output file must begin with two special comment lines specifying the experiment run and inputs used. These lines must appear at the beginning of the file and are to be formatted as follows:

The first line is a special comment specifying the experiment as defined in Section 7.3.1 (EXP-ID) and is of the form:

;; EXP-ID: <EXP-ID> [21]

For example,

;;EXP-ID: sri_03_edit_eval03_eng_bnews_ref_1

The second line is a special comment specifying the inputs used (excluding speech files) and is of the form: [25]

;;INPUTS: <FILE1>,<FILE2>, ... <FILEn>

For example,

;;INPUTS: sw_47620.ctm

Note: Minimally, the INPUTS must specify a CTM file containing the token inputs used by the MDE system (if it used a word/token input source). The referenced CTM file may exist in any directory in the submission by preceding the CTM filename with the relative path. If no path information is included, a copy of the CTM file must be included in the same directory as the MDTM file referencing it. This information will be used in token-alignment-based scoring.

The header is followed by a list of output records. The output record format is as follows:

<SOURCE><SP><CHANNEL><SP><BEG-TIME><SP>
<DURATION><SP><TYPE><SP><CONF><SP>
<SUBTYPE><SP><SPEAKER> <NEW-LINE>

where,

<SP> is whitespace.

<SOURCE> is the basename of the source audio file being processed (no pathnames or extensions should be included)

<CHANNEL> is an integer that specifies the source channel which exhibited the metadata event. The first channel is "1", the second is "2", and so on. This value will always be "1" for single-channel files.

<BEG-TIME > is the beginning time of the metadata event. This time is a floating point number, expressed in seconds, measured from the start time of the file. [22]

<DURATION> is the duration of the metadata event. This time is a floating point number, expressed in seconds. [22]

<TYPE> is a string specifying the type of metadata event being output. Permitted TYPE values are:

filler | edit | su | speaker

(Note: only the "speaker" type and subtypes will be used in RT-03S)

<CONF> is a decimal value between 0 and 1 expressing the system's confidence in the existence of the metadata event being output. A value of "NA" is to be used when no confidence is computed and in the reference data. [26]

<SUBTYPE> is a string specifying the subtype of the metadata event being output.

[24] Any token which is to be excluded from scoring may be given this tag – including those for which specified types exist. However, where possible, sites are encouraged to use the supported types to enhance the usefulness of the data for MDE experiments.

[25] An INPUTS entry must be made for any ancillary CTM input used, including the NIST reference transcript. The ancillary CTM file must also be included – even if it was the NIST reference. This is because some sites alter the reference transcription using forced alignment. This information will be used in word-based-alignment scoring.

[26] MDE systems are not required to emit a confidence score and may use the "NA" value in all output records.

For TYPE = “filler”, the legal SUBTYPE values are:

filled_pause | discourse_marker | explicit_editing_term

For TYPE = “edit”, the legal SUBTYPE values are:

none | repetition | revision | restart | complex[27]

For TYPE = “su”, the legal SUBTYPE values are:

statement| question | backchannel | incomplete

For TYPE = “speaker”, the legal SUBTYPE values are:

adult_male | adult_female | child | unknown

<SPEAKER> is an ASCII character string that uniquely identifies the speaker. This field is optional except for the speaker diarization task.

Here are some example MDE output records:

```
file52 1 1234.56 2.23 su 0.72 question
file03 2 52.77 1.33 disfl 0 none 2
```

7.2.4 SYSTEM DESCRIPTION

For each test you run (for each unique EXP-ID), a brief description of the system (algorithms, data, configuration) used to produce your system output must be provided along with your system output. If you submit multiple system runs for a particular experiment with different systems/configurations, you must explicitly designate one run as the primary system and the others as contrastive systems in your system description. This information is to be recorded in a file named:

<EXP-ID>.txt

(where EXP-ID is defined in Section 7.3.1)

and placed in your “output” directory along side the similarly-named directories containing your system output. This file is to be formatted as follows:

1. EXP-ID = <EXP-ID>

2. Primary: yes | no

3. System Description:

   [brief technical description of your system; if a contrastive test, contrast with primary system description]

4. Training:

   [list of resources used for training; for STT, be sure to address acoustic and LM training, and lexicon]

5. Execution Time (STT only):

   [sites must report the time that was required to process the test data as described in Section 6.1; sites should include a description of the CPU and amount of memory used]

6. References:

   [any pertinent references]

7.3 SUBMISSION INSTRUCTIONS

7.3.1 SUBMISSION EXPERIMENT CODES

The output of each submitted experiment must be identified by the following code as specified above.

EXP-ID =
<SITE>_<YEAR>_<TASK>_<DATA>_<LANG>_
<TYPE>_<COND >_<SYSID>_<RUN>

Where,

SITE ::= expt | bbn | bbnplus | cu | elisa | clips | sri | sriplus | ibm | mitll | ms | pan | ...

(The special SITE code “expt” is used in the EXP-ID-based filename of the UEM test index files under the “indices/” directory to list the test material for a particular experiment and in the EXP-ID-based subdirectory name under the “input/” directory to indicate ancillary data to be used in certain control condition experiments.)

YEAR ::= 03

For the RT-03 Evaluation, these are:

TASK ::= stt1x | stt10x | sttul | stt1xmb | stt10xmb | sttulmb | spkr | data

where,

stt1x = STT run at 1X realtime

stt10x = STT run at 10X realtime

sttul = STT run with “unlimited” processing time

stt1xmb = STT mothballed RT-02 system run at 1X realtime

stt10xmb = STT mothballed RT-02 system run at 10x realtime

sttulmb = STT mothballed RT-02 system run with “unlimited” processing time

spkr = speaker diarization

data = a special TASK code be used to provide a directory for ancillary data such as common CTM files used over many MDE experiments. Please make sure to use increasing run numbers for this special experiment ID when making multiple submissions so that your ancillary data from earlier submissions is not over-written here at NIST

DATA ::= eval03 | prog

Where,

eval03 = all “Current” data sets (English, Mandarin, Arabic)

prog = Progress Test Set

(note that the MDE data is not listed explicitly as a separate dataset since it is a proper subset of eval03)

LANG ::= eng | arab | mand

[27] The subtype is an optional field for EVAL-TYPE=edit in EARS 2003. Sites wishing not to output a subtype in this case should output "none" as the EVAL-SUBTYPE value.

TYPE ::= bnews | cts

CONDITION ::= spch | ref

Where,

spch = audio input only

ref = audio input + reference transcript input

(Although the "spch" [speech] condition is the primary condition of interest, the "ref" [reference] condition is provided so as to provide a control for speech recognition (if an STT system was used in performing MDE) and includes both the speech and reference transcript as input. It is not applicable for the "stt*" tasks. The "spkr" task for this condition may make use of only the lexical tokens in the reference transcript [no reference speaker segmentation may be used.])

SYSID ::= site-named string designating the system used

[This is intended so that we can differentiate between contrastive runs for the same condition. Therefore, a different SYSID should be created for runs where any manual changes were made to a particular system]

RUN ::= 1..n (with values greater than 1 indicating multiple runs of the same experiment/system)

[An incremental run number MUST be used for multiple submissions of any particular experiment with an identical configuration (due to a bug or runtime problem.) This should NOT be used to indicate contrastive runs. Instead, a different SYSID should be used. However, please note that ONLY the first run will be considered "official" and will be scored by NIST unless special arrangements are made with NIST. Please also note that submissions which reuse identical experiment IDs/run numbers from previous submissions will be automatically rejected.]

examples:

bbn_03_stt10x_eval03_eng_cts_spch_superreco1_1

sri_03_spkr_eval03_eng_bnews_ref_speakerid2_1

7.3.2 SUBMISSION DIRECTORY STRUCTURE

All system output submissions must be formatted according to the following directory structure:

output/<SYSTEM-DESCRIPTION-FILES>

(one for each EXP-ID as specified in 7.2.4)

output/<EXP-ID>/ <OUTPUT-FILES>

where,

<EXP-ID> is as defined in Section 7.3.1

<OUTPUT-FILES> are as defined in Section 7.2

Note: one output file must be generated for EACH input file as specified in the test index for the experiment being run. The output files are to be named so as to be identical to the input file basenames with the appropriate .ctm or .mdtm filetype extension. For example, an STT output file for the speech waveform file sw_47620.sph must be named sw_47620.ctm and an MDE output file must be named sw_47620.mdtm. When generated, these output files are to be placed under the appropriately-named EXP-ID directory on your system identifying the experiment run.

7.3.3 SUBMISSION PACKAGING AND UPLOADING

To prepare your submission, first create the previously-described file/directory structure. This structure may contain the output of multiple experiments, although you are free to submit one experiment at a time if you like. The following instructions assume that you are using the UNIX operating system. If you do not have access to UNIX utilities or ftp, please contact NIST to make alternate arrangements.

First change directory to the parent directory of your "output/" directory. Next, type the following command:

```
tar -cvf - ./output | gzip > <SITE>_<SUB-NUM>.tgz
```

where,

<SITE> is the ID for your site as given in Section 7.3.1

<SUB-NUM> is an integer 1 – n where 1 identifies your first submission, 2 your second, and so forth.

This command creates a single tar file containing all of your results. Next, ftp to jaguar.ncsl.nist.gov giving the username 'anonymous' and your e-mail address as the password. After you are logged in, issue the following set of commands, (the prompt will be 'ftp>'):

```
ftp> cd incoming
ftp> binary
ftp> put <SITE>_<SUB-NUM>.tgz
ftp> quit
```

You've now submitted your recognition results to NIST. The last thing you need to do is send an e-mail message to Audrey Le at audrey.le@nist.gov to notify NIST of your submission. The following information should be included in your email:

1) The name of your submission file

2) A listing of each of your submitted experiment IDs

3) e.g.,

```
Submission: bbnplus_1 <NL>
Experiments: <NL>
bbnplus_03_stt10x_eval03_eng_cts_spch
_superreco1_1<NL>
bbnplus_03_stt10x_eval03_eng_cts_spch
_superreco2_1 <NL>
```

.

.

Note that submissions received after the stated due dates FOR ANY REASON will be marked late. So, please submit your files in time for us to deal with any transmission/formatting problems that might occur well before the due date if possible.

8 SCHEDULE

The evaluation schedule below is accurate at the time this document was published. Please consult the live version of the schedule at http://macears.ll.mit.edu/macears_docs/ears-schedule.txt for any late-breaking changes.

03 Mar - NIST releases English and Non-English Current test data

(NIST will release baseline segmentations created by MIT-LL as soon as possible thereafter)

(Sites freeze English STT systems and begin tests sometime before the submission due date [28])

07 Apr - Sites submit STT outputs from English portion of Current Test Set

08 Apr - NIST releases English Current Test Set STT outputs plus reference data (transcripts and segmentation) needed for the MDE evaluation

09 Apr - NIST releases Progress test data to sites

14 Apr - NIST releases STT scores for English portion of Current Test Set

22 Apr - Sites submit MDE outputs from English portion of Current Test Set

23 Apr - Sites submit STT outputs from Chinese and Arabic portions of Current Test Set

23 Apr - Sites submit STT outputs from RT-03 systems and mothballed RT-02 systems on Progress Test Set

29 Apr - NIST releases MDE scores for English portion of Current Test Set

30 Apr - NIST releases STT scores for Chinese and Arabic portions of Current Test Set

30 Apr - NIST releases STT scores for mothballed RT-02 systems and RT-03 systems on Progress Test Set

14 May - Slides for RT-03 and EARS workshop notebooks due

19-20 May - RT-03S Workshop

21-22 May - EARS PI Meeting

Please note that the stated dates are hard deadlines. All late submissions will be marked as such and given the tight schedule, severely late submissions may not be scored at all prior to the workshops.

9 WORKSHOPS

The evaluation will be followed by two back-to-back workshops. The first, the Rich Transcription 2003 Spring (RT-03S) Workshop, is open to all participants. The RT-03 Workshop is immediately followed by an EARS PI meeting which is only open to EARS contractors and affiliates. Information regarding workshop logistics and registration will be posted at a later date in email.

[28] Each site may select its own date for freezing its system. The point of this entry is to reinforce the fact that each site must freeze its English STT system (sometime in the 3 Mar - 7 Apr range) before beginning the English Current test -- and keep the system frozen through the Progress test.

Table 1. Composition of the EARS 2003 evaluation test set

Use	Domain	English STT (Progress Set)	English STT (Current Set)	English MDE	Chinese STT	Arabic STT

Evaluation	Broadcast News	180 minutes: 6 30-minute excerpts from TDT-4: 6 sources/1 show per source/first 30 minutes transcribed to story boundary, source TDT4, from February 2001.	180 minutes: 6 30-minute excerpts from TDT-4: 6 sources/1 show per source/first 30 minutes transcribed to story boundary, source TDT4, from February 2001.	Spring
				90 minutes: diarization using first 3 (chronologically) of the 6 STT Current shows;
				Fall
				90 minutes: all MDE tasks using last 3 (chronologically) of the 6 STT Current shows;
					60 minutes: 12 minutes from 5 TDT-4 sources, from February 2001	60 minutes: 30 minutes from 2 TDT-4 sources, from February 2001

	Conversational Telephone	180 minutes: 36 5-minute conversation excerpts: conversation sides approx. 18 cellular/54 landline, 36 male/36 female, balanced by age, dialect region, and topic, 5 consecutive transcribed minutes chosen to be topically oriented without major noise or interference and delimited by turn boundaries, source is Fisher-2003	360 minutes: 72 5-minute conversation excerpts: 36 Fisher-2003 chosen and transcribed similarly to the Progress Set and 36 selected (balanced) SWBD-Cell[29] with minutes 1-6 (starting and ending on turn boundaries) transcribed	Spring 90 minutes: Diarization using 18 of the STT Current excerpts, balanced by source (SWBD-Cell[29] /Fisher) and other characteristics; Fall 90 minutes: Diarization using another 18 of the STT Current excerpts, with same balance as the Spring set;	60 minutes: 5 minutes from 12 unused CallFriend Mandarin data	60 minutes: 5 minutes from 12 unused CallHome Egyptian Arabic data
Development	Broadcast News	NA	RT-02 60-minute evaluation data (6 10-minute sources), from December 14-19, 1998 No data from February 2001 (the test epoch) or later may be used	90 minutes: each of the 3 shows used in Spring diarization evaluation	None no data from February 2001 or later (the test epoch) may be used	None no data from February 2001 or later (the test epoch) may be used
	Conversational Telephone	NA	RT-02 300-minute evaluation data (5 minutes from 20 SWBD-1 + 20 SWBD-2.2 +20 SWBD-Cel1)	90 minutes: each of the 18 excerpts used in the Spring diarization evaluation	None	None
Training	Broadcast News	NA	All previously released BN data + 180 minutes new data from TDT-4: 6 30-minute excerpts, 6 sources, 1 show per source, first 30 minutes transcribed; no data from February 2001 (the test epoch) or later may be used	All previously released STT BN training data may be usable for diarization, data for other MDE tasks pending	All previously released Mandarin BN data; no data from February 2001 or later (the test epoch) may be used	All previously released Arabic BN data; no data from February 2001 or later (the test epoch) may be used
	Conversational Telephone	NA	All previously released CTS data	All previously released STT CTS training data may be usable for diarization, data for other MDE tasks pending	All previously released Mandarin CTS data	All previously released Arabic CTS data

[29] Although the source is SWBD-Cellular, the calls include a balance of land-land, land-cell, and cell-cell calls