NIST Rich Transcription 2002 Evaluation: A Preview

John Garofolo, Jonathan G. Fiscus,
Alvin Martin, David Pallett, Mark Przybocki

National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8940
Gaithersburg, MD 20899-8940, USA
{jgarofolo@nist.gov, jfiscus@nist.gov, amartin@nist.gov,
dpallett@nist.gov, mprzybocki@nist.gov}

Abstract

The National Institute of Standards and Technology (NIST) has been implementing evaluations of automatic speech transcription technologies for over 15 years. NIST has helped guide progress in these technologies by: creating increasingly challenging and realistic tests, helping to provide associated linguistic resources, employing uniform metrics and analyses across systems to assess performance, and sponsoring evaluation-related technology workshops. Over time, this approach has shown great progress in the technology as the test domains have become more difficult and error rates have almost consistently decreased. In conjunction with the new DARPA Effective, Affordable, Reusable Speech (EARS) Program, NIST has begun an evaluation effort to help move the state-of-the-art to the next level in the form of a *Rich Transcription (RT)* evaluation program. RT is defined to be an integrated combination of speech-to-text generation (STT) and metadata (MD) annotation as applied to multiple domains such as speech from Broadcast News, telephone conversations, and meetings. The Rich Transcription 2002 (RT-02) evaluation will have been the first in an annual series of evaluations and workshops focusing on this technology.

1. Introduction

For over fifteen years, NIST has been conducting common evaluations of the performance of automatic speech recognition technology. Traditionally, these evaluations have focused on the accuracy of automatically-generated orthographic word transcriptions. Over the years, as the technology improved, these evaluations evolved from transcription of contrived limited-vocabulary scripts read by subjects in sound isolation booths to transcription of real news broadcasts and telephone conversations. However, the core approaches used today remain relatively the same as they did ten years ago. Many of the recent improvements in the technology are largely due to faster processor speeds, cheaper memory, and the availability of large, specialized training corpora.

It is widely recognized that it is time for a shift in the approaches used in automatic speech recognition toward utilizing increased knowledge regarding language, context, and world knowledge and the integration of Speech-to-Text (STT) technology with other recognition/language processing technologies. It is also believed that the addition of feedback from these integrated technologies will improve the core STT performance itself in synergistic ways. Such changes will help make the output of recognition systems more accurate and more useful for a variety of applications. Further, many believe the time has also come to focus on building more generic ASR capabilities that can be ported to new domains and vocabularies without necessitating complete rebuilds of systems using huge quantities of domain-specific training and language modeling data.

It is clear that the technology is ready for these evolutionary changes. The fledgling DARPA Effective Affordable Reusable Speech (EARS) program is setting out to act as a catalyst for such change via four thrusts: 1) improvement of the core technology incorporating tight integration of metadata extraction to provide both human- and machine-usable transcripts; 2) exploration of novel approaches including improved auditory modeling, the use of prosodic information, dynamic language models, and more; 3) supplying the necessary linguistic resources to support development and evaluation; and 4) developing and evaluating usable interfaces for the output of these systems in the context of integrated user-based systems.[EARS BAA, 2002]

This paper will focus on NIST's approach to support for the development and evaluation of the core STT technology as integrated with automatic metadata (MD) annotation. The product of such technology is referred to as *Rich Transcription* (RT). [1]

The Rich Transcription 2002 Workshop (RT-02) was the first in a series of evaluations that will help to propel the technology forward. The RT-02 evaluation took place in April 2002 and was followed by a workshop in May. This paper describes the RT concept, discusses the RT-02 evaluation plan, and briefly takes a look toward the future. Since this paper was submitted prior to the conclusion of the evaluation, no results are given. The results of the evaluation will be published in the proceedings of the RT-02 Workshop.

2. Rich Transcription

Automatic speech recognition systems of the near future will be required to output a variety of metadata integrated with orthography. These enriched transcripts will be more

[1] In an abstract sense, of course, word transcription itself can be viewed as a kind of metadata annotation where words are abstract symbols referencing semantic elements in the speech signal. However, for the purposes of this discussion, speech-to-text generation and metadata annotation will be treated as separate forms of technology.[1]

useful for downstream processing by search, extraction, summarization, and translation technologies, as well as provide the necessary information to create a version of the transcripts with capitalization, punctuation, and other formatting for greatly improved human readability. We refer to these enriched systems as Rich Transcription (RT) systems.

To support research and evaluation in automatic RT, NIST conducted a pilot evaluation, Rich Transcription 2002 (RT-02) which, in addition to the evaluation of orthographic transcription, also addressed the automatic production of metadata.[RT-02, 2002]

Figure 1 shows an example of an old-style ASR transcript enriched with metadata using an XML construct. It is clear that the product of this enrichment, which can be readily translated into a human-readable form, is vastly superior to the unsegmented, uncapitalized, and unpunctuated word-stream transcripts produced by traditional ASR systems.

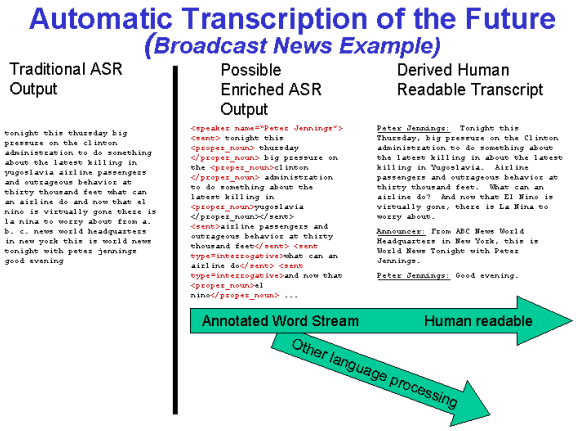

Figure 1: Rich Transcription Example

However, it is conceivable that a great deal more metadata than is shown in the example could be applied to the transcript to make it even more usable for particular applications. The exploration of new metadata types will be an integral part of the RT evaluation series. As such, the supported metadata types are likely to change/evolve over time.

3. Initial Metadata Experiments at NIST

Over the last year and a half, NIST has conducted three informal experiments to explore the application of metadata annotation to automatic speech recognition. The first two experiments, which were completely internal to NIST, involved the annotation of a data set by multiple annotators to explore the problem space and gauge inter-annotator consistency. The third experiment solicited input from the community.

In the Fall of 2000, we ran an open experiment on the Internet to quantify the kinds of output researchers would expect from the next generation of speech recognition systems. Participants in the experiment were given a sample of Broadcast News audio and asked to generate a hypothetical ASR transcript of the future. The instructions for the experiment were simple, and purposefully vague: "render the broadcast news excerpts in a manner that you find easily readable." The participants were given only the recordings so as to not bias them towards existing transcription standards.

The results were both encouraging and surprisingly uniform given the vague instructions. All respondents included the following types of markup: speaker identification, sentence boundaries, capitalization, numeric normalizations and punctuation. The results were particularly surprising because there was considerable inter-respondent agreement for both manner and type of markup for all the formatting types except comma placement (which is known to be largely subjective.) As a result of this experiment, NIST began developing the concept of enriched transcripts.

In December 2001, in preparation for our first Rich Transcription evaluation, NIST set out to explore the feasibility of annotating several types of speech with what was thought at first to be a relatively simple set of metadata. These metadata included speaker changes and identity and sentence boundaries and type. The sentence types chosen were *interrogative*, *declarative*, *exclamatory*, and *imperative*. The goal of the experiment was to see if humans could consistently and reliably produce these annotations and to produce sample data for researchers to begin developing RT systems.

Recordings from three domains were chosen: news broadcasts, telephone conversations, and multi-participant meetings. NIST annotators were assigned data sets so that each data set was annotated by at least three different people.

The speaker segmentation and identification process was relatively straightforward. However, the sentence annotation proved to be much more problematic than originally anticipated. This is largely because much of our prior work had been with Broadcast News data that contains a great deal of scripted/non-spontaneous speech. When we began working with conversational/interactive speech, we learned that it was extremely difficult to achieve any degree of inter-annotator consistency in annotating “sentences”. Further, individual annotators had difficulty formulating their own criteria for marking sentence boundaries within the spontaneous data.

These results led us to rethink our original goals for the near term. We decided to run a community-wide experiment to solicit input from researchers on useful metadata annotation types. The researchers were instructed to propose metadata annotation types that could be annotated consistently and simply and for which automatic annotation systems with some degree of accuracy could be created in the relative near term.

We created a sample data set with three short excerpts each from news broadcasts, telephone conversations, and meetings. We instructed the experimenters to provide a definition for each of their proposed metadata types and annotate the sample data. To simplify the process for the researchers and to focus the experiment, we pre-

transcribed the data and provided speaker segmentation which we asked the experimenters to use. We received results from IBM, ICSI, LIMSI, and MIT Lincoln Labs. The following are a summary of the metadata types proposed in the experiment broken down into non-linguistically-motivated and linguistically-motivated categories:

Non-Linguistically-Motivated Metadata:

- speaker gender
- multiple simultaneous sound-producing sources
- bandwidth
- music
- noise (vocal and non-vocal)

Linguistically-Motivated Metadata:

- punctuation, capitalization, and formatting
- named entity and type
- utterance boundary and type
- disruption point
- verbal edit interval
- filled pause
- quotation
- parenthetical/aside

The raw results of the experiment are available on the RT-02 website. [MD experiment, 2002]. We continue to welcome suggestions for new metadata types within our experimental framework.

After the conclusion of the final experiment, we tentatively settled on the following metadata types for inclusion in the RT-02 evaluation:

- Speaker segmentation and identification
- Sentence or phrasal unit segmentation and classification
- Acronym detection and expansion
- Verbal edit detection, identifying regions of disfluency
- Named entity detection/classification
- Numeric expression detection/classification
- Temporal expression detection/classification

With the exception of sentence segmentation, we believed that these types could be implemented (albeit some with high error rates) with current technology and would be very useful toward the goal of producing human-readable transcripts.

However, given our extremely compressed schedule, we chose to implement a proof-of-concept evaluation that would permit us to begin to build the infrastructure for RT evaluation while not making undue demands on the organizers or the participants. We settled on an evaluation that would stress systems with regard to the speech transcription tasks and which would include a speaker segmentation task as a placeholder for metadata annotation.

4. Rich Transcription 2002 Evaluation

The RT-02 evaluation plan specified no required test conditions since this was a pilot evaluation that was open-ended by design to encourage participation and experimentation. However, several conditions were *suggested* in the evaluation plan so that the evaluation would produce a baseline for conditions which are likely to be of interest in future evaluations. [RT-02 Evaluation Plan, 2002]

As described above, Rich Transcription, is defined to contain two primary types of language technologies: Speech-to-Text (STT) transcription and Metadata (MD) annotation. As in similar earlier composite technology evaluations, such as the TREC Spoken Document Retrieval Task [Garofolo et al., 2000] and the Hub-4 Entity Recognition Task, [Burger et al., 1998 and Przybocki et al., 1999] the evaluation was designed to examine the performance of the individual component technologies. To maximize participation this year, participants were permitted to run only the STT or MD task or work with only the test material from one or two of the three supported domains. However, in order for us to obtain an informative baseline for current capabilities, we encouraged participants to run as many supported tasks and conditions as possible – even if poor results were expected. Further, participants were encouraged to team up with other participants with complementary capabilities to create full RT systems. This mode of participation is likely to become even more prevalent in future RT evaluations in which a wide variety of metadata annotation tasks are likely to be supported.

4.1. Test Corpora

The RT-02 evaluation data set contained material from three distinct domains: 1) Broadcast News, 2) Telephone Conversations, and 3) Meetings as follows:

4.1.1. Broadcast News Subset

This subset consisted of approximately 1 hour of television and radio news broadcast excerpts taken from previously unreleased LDC Broadcast News Corpora. [LDC, 2002] Unlike previous Hub-4 BN tests, no concatenation of excerpts was performed. Rather, whole broadcasts were made available and an index specifying excerpts for evaluation was provided. Test participants were permitted to use the non-evaluation material for unsupervised adaptation within a broadcast, however cross-broadcast adaptation was not permitted.

4.1.2. Telephone Conversation Subset

This subset consisted of 300 minutes of recordings of 2-channel (one for each participant in the conversation) telephone conversations and is similar to that used for the 2001 Hub-5 evaluation [Hub-5, 2001]. The test material included three subsets: 1) unreleased original Switchboard Corpus material, 2) Switchboard II Phase 3 Corpus material, and 3) Switchboard Cellular Phase 2 Corpus material. [LDC, 2002] Each of the subsets contained 5 minutes from each of 20 conversations.

4.1.3. Meeting Subset

This subset consisted of approximately 80 minutes of excerpts of meetings recorded at CMU, ICSI, the LDC, and NIST. As such, the represented speakers, forums, vocabularies, recording conditions, and recording equipment were quite diverse. Two complete meetings were provided from each site and an index specifying a

10-minute excerpt within each meeting for evaluation was provided. As in the BN data, participants were permitted to perform unsupervised adaptation within a meeting, but cross-meeting adaptation was not permitted. The meetings contained speech from 3 to 8 participants. Each meeting was represented by 3 forms of recording: 1) a recording from an omni-directional centrally-placed microphone, 2) a recording from each of a personal microphone placed on each participant (either a head-mounted boom microphone or a lapel/lavalier microphone), and 3) a gain-adjusted mix of the personal microphones. The meetings contained both native and non-native speakers of English.

4.2. Training Corpora

Any material publicly available at the time of the start of the evaluation could be used for training. Following previous Hub-4 conventions, for the Broadcast News subtest, only news-based material dated prior to December 31, 1998 could be used for training purposes.

Since no publicly available training corpus for meeting recognition existed at the time of the evaluation, a small data set with similar size and properties to the evaluation test set was provided for training. It is understood that such a set is far too small for automated training purposes. Rather, this set could be used for developmental testing or manual training purposes at the discretion of the test participant.

4.3. Development Test Corpora

No specific development test corpora were made available for this evaluation.

4.4. Evaluation Tasks and Conditions

This evaluation supported two primary evaluation tasks and several conditions within each task.

4.4.1. Speech-to-Text (STT) Generation Task

This task was similar in nature to previous NIST ASR transcription tasks with the basic goal of generating a word stream from speech input. To simplify implementation and permit backward compatibility, participants were required to provide their output in conformance with previous Hub-5 transcription conventions. Although participants were permitted to perform unsupervised adaptation within a recorded event (news broadcast, telephone conversation, meeting), participants were not permitted to perform cross-event adaptation. Note, however, that participants were permitted to make use of multiple recordings of an event (for telephone conversations and meetings where multiple channels were collected) for certain conditions.

4.4.2. Metadata (MD) Annotation Task

The metadata annotation task for this evaluation consisted of segmenting audio excerpts into speaker changes and clustering these segments by speaker. As with the STT task, within-event unsupervised adaptation was permitted, but cross-event adaptation was forbidden.

4.4.3. Evaluation Conditions

Processing Speed:
Participants were required to provide specific information about the processing speed for the task implemented and categorize each run as being either : 1) greater than 10-times realtime, 2) less-than-or-equal-to 10 times realtime, or 3) realtime or faster.

Domain:
This evaluation included material from three distinct domains (Broadcast News, telephone conversations, and meetings). Test participants were permitted to choose a subset of the domains for evaluation at their discretion. However, participants were encouraged to implement their systems on all of the data to provide a baseline for future work.

Channels:
As was described above, certain domain sets included multiple event recordings (i.e., two channels for each phone conversations and different microphones/mixes for meetings). Participants were encouraged to implement runs on each channel set. For the meeting data, the omni-directional microphone channel would provide a realistic "high bar", whereas the close-talking microphone mix channel would provide a high quality signal version of the same data, and the individual close-talking microphone channels would provide a control for speaker overlap.

Segmentation:
Manual speaker segmentation was provided for participants who wished to implement the traditional Hub-5 LVCSR evaluation condition where manual speaker segmentation is given. Manual speaker segmentation was also provided for the meeting data as a contrast condition. The output of the CMU Hub-4 segmenter on the BN data was also made available for participants without access to their own segmenter. Unfortunately, no such segmenter was available to NIST for the telephone or meeting domain data. In future evaluations, NIST would like to provide automatic segmentation for all three domains.

4.5. Scoring and Evaluation

The Speech-to-Text generation task was evaluated using the NIST SCLITE speech recognition scoring software and, as in past such evaluations, Word Error Rate was the primary metric. [RT-02 Scoring, 2002] Unlike in past evaluations, however, areas of overlapping speech were evaluated. The rules for orthography generation were provided in the RT-02 Evaluation plan. [RT-02 Evaluation Plan, 2002]

The Metadata annotation task was evaluated using the NIST speaker segmentation scoring software. Speaker segmentation systems were evaluated using the total segmentation error metric as defined for the Speaker Segmentation task in the 2001 Speaker Recognition Evaluation Plan [Speaker Recognition, 2001]. Total segmentation error is essentially 1 minus the ratio between the amount of time correctly segmented by the system and the amount of time speakers were talking. The complete algorithm is given in the RT-02 Evaluation Plan. [RT-02 Evaluation Plan, 2002] However, unlike in past

segmentation evaluations, areas of overlapping speech were evaluated. [RT-02 Scoring, 2002]

5. Future RT Evaluations

Since this was a pilot evaluation of RT technology and since it was implemented under very tight time constraints, it employed existing infrastructure where possible and addressed RT in a very limited way. It is therefore expected that future such evaluations will have very different characteristics.

Firstly, it is likely that future RT evaluations will employ comprehensive formats, representations, and evaluation software that will support the integrative goals of the program and that can evolve as needed. Toward this end, we are currently working toward creating a generic evaluation engine based on the ATLAS architecture that can be used for a variety of recognition, detection, and classification tasks through the use of an evaluation-task-based configuration mechanism. [ATLAS, 2002 and Laprun, et al., 2002]

Secondly, future RT evaluations are likely to support a variety of metadata types such as those suggested earlier. A variety of MD types have been suggested in the experiments already run and it is likely that more will come about as the program evolves. The MD types to be addressed in the next RT evaluation will be the result of further experiments and community input during and after the RT-02 Workshop. It is therefore likely that the RT metadata types will evolve over time.

Thirdly, it is likely that a set of required evaluation tasks and conditions will be employed to focus the research and permit comparisons across systems. Further, since cross-test year-to-year comparisons provide a valuable tool for both research sites and sponsors, the test sets and scoring paradigms for the STT task will be tuned to permit such comparisons.

6. References

ATLAS (2002), Architecture and Tools for Linguistic Annotation Systems website: http://www.nist.gov/speech/atlas/

Burger, J., Palmer, D., and Hirschman, L (1998), *Named Entity Scoring for Speech Input*, Proc 36'th Annual Meeting of the Association for Computational Linguistics (ACL/COLING '98), August 1998.

EARS BAA (2002), DARPA EARS Program CBD Broad AgencyAnnouncement http://www.darpa.mil/ito/solicitations/cbd_02-06.html

Garofolo, J., Auzanne, C., and Voorhees, E., (2000), *The TREC Spoken Document Retrieval Track: A Success Story*, Proc. RIAO'2000, Vol 1. pp 1-20.

Hub-5 (2001), Hub-5 2001 Website http://www.nist.gov/speech/tests/ctr/h5_2001/index.htm

Laprun, C., Fiscus, J., Garofolo, J., and Pajot, S. (2002), *A Practical Introduction to ATLAS*, Proc. LREC-2002.

LDC (2002), Broadcast News and conversational telephone corpora used in this test were provided by the Linguistic Data Consortium: http://www.ldc.upenn.edu

RT-02 (2002), NIST RT-02 Website http://www.nist.gov/speech/tests/rt/rt2002/

MD Experiment (2002), RT-02 Metadata Annotation Experment http://www.nist.gov/speech/tests/rt/rt2002/experiment.htm

Przybocki, M., Fiscus, J., Garofolo., J., and Pallett, D., (1999), *1998 Hub-4 Information Extraction Evaluation*, Proc. 1999 DARPA Broadcast News Workshop: http://www.nist.gov/speech/publications/darpa99/html/ov20/ov20.htm

RT-02 Evaluation Plan (2002), NIST Rich Transcription 2002 Evaluation Plan http://www.nist.gov/speech/tests/rt/rt2002/rt02_eval_plan_v1.pdf

RT-02 Scoring (2002), RT-02 Evaluation Software http://www.nist.gov/speech/tests/rt/rt2002/evalsoftware.htm

Speaker Recognition (2001), The NIST Year 2001 Speaker Recognition Evaluation Plan http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf