The NIST Year 2003 Speaker Recognition Evaluation Plan

1 INTRODUCTION

The year 2003 speaker recognition evaluation is part of an ongoing series of yearly evaluations conducted by NIST. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end the evaluation is designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to those wishing to participate.

The evaluation will be conducted in the spring. The data will be made available to participants in April, with results due to be submitted to NIST about four weeks later. A follow-up workshop for evaluation participants to discuss research findings will be held in June. Specific dates are listed in section 11, Schedule.

Participation in the evaluation is invited for all sites that find the tasks and the evaluation of interest. For more information, and to register to participate in the evaluation, please contact Dr. Alvin Martin at NIST.[1]

2 TECHNICAL OBJECTIVE

This speaker recognition evaluation focuses on the task of speaker detection. This task is posed primarily in the context of conversational telephone speech. The evaluation is designed to foster research progress, with the goals of:

- Exploring promising new ideas in speaker recognition.
- Developing advanced technology incorporating these ideas.
- Measuring the performance of this technology.

2.1 Task Definition

The year 2003 speaker recognition evaluation plan is limited the following broadly defined task[2]:

2.1.1 Speaker detection

This task is NIST's basic speaker recognition task. The task is to determine whether a specified speaker is speaking during a given speech segment.[3]

2.2 Task Conditions

The year 2003 speaker recognition evaluation plan includes three distinct task conditions for the speaker detection task. These are as follows:

2.2.1 One-speaker detection – limited data

The one-speaker detection task condition will remain essentially the same as in previous years. As in 2002, the data will be taken from the second release of the cellular switchboard corpus (Switchboard Cellular – Part 2) of the Linguistic Data Consortium (LDC).[4] The training data for a target speaker will be two minutes of speech from that speaker, excerpted from a single conversation. Each test segment will be the speech of a single speaker, excerpted from a one-minute segment taken from a (single) conversation.

While the underlying data will be recycled from the 2002 evaluation, the selection of which conversations will be used for training and which for test, and of the specific speech data from conversation sides that compose each training or test segment will be different from that of last year. (There will also be some trials using two-speaker training data and one-speaker test segments. See section 4.1.2.)

2.2.2 Two-speaker detection – limited data

The two-speaker detection task condition differs from the one-speaker conditions in that no excerpting of the speech of a single speaker is performed, neither in training nor in test. The data will be taken from the same (recycled) source as that of the one-speaker detection condition with limited data. Thus each test segment will be a one-minute segment taken from a (single) conversation, but no excerpting will be performed and the two sides of the conversation will be summed together. Further, the training data for a target speaker will be three whole conversations, with the two sides of the conversation summed together. Thus a major challenge in training will be to discover which of the two speakers in each of the training conversations is the target speaker.[5] (There will also be some trials using one-speaker training data and two-speaker test segments. See section 4.1.2.)

2.2.3 One-speaker detection – extended data

This task condition provides a much larger amount of training data for target speakers – up to an hour of speech per speaker. The intent is to foster new research on improving speaker recognition performance through the discovery and exploitation of higher-level and more complex characteristics of a speaker's speech, such as idiosyncratic language patterns and nonlinguistic vocalizations.

---

[1] To contact Dr. Martin, send him email at alvin.martin@nist.gov, or call him at 301/975-3169.

[2] The speaker segmentation task that was supported in previous NIST speaker recognition evaluations is now being included in the NIST coordinated EARS evaluations. The EARS 2003 evaluation, to be held in the spring prior to the speaker recognition evaluation, includes (speaker) diarization as one of its metadata extraction subtasks. This subtask is open to interested outside sites either as individual participants or in a team with sites participating in other aspects of the evaluation. For further information on this evaluation, see http://www.nist.gov/speech/tests/rt/rt2003/

[3] In previous evaluation plans, the speaker detection task was divided into a "one-speaker" and a "two-speaker" task. However, this distinction relates to the task conditions rather than the task definition. Therefore in this evaluation plan the one- and two-speaker conditions have been moved to section 2.2, task conditions.

[4] Refer to www.ldc.upenn.edu/Projects/.

[5] The non-target speakers appearing in a speaker's training data will be controlled so that no non-target speaker appears in more than one training conversation.

The data will be taken from LDC's switchboard II corpus, phases 2 and 3. This is identical to the data used last year, so this condition will basically involve a repetition of the test in the 2002 evaluation. However, more auxiliary types of data will be made optionally available to participating sites that in 2002. See section 4.1.4. Recent research suggests significant improvement of performance results beyond those of the 2002 evaluation may be attainable.

The training data for a speaker will be all of a target speaker's speech from N whole conversations, with N varying between 1 and 16. Each segment of test segment data will be all of one side of a single conversation. (not both sides summed together)

3 PERFORMANCE MEASURES

Evaluation will be performed separately for each of the tasks and task conditions in section 2. Evaluation of all these tasks and task conditions will use cost-based performance measures. A cost-based performance measure is used so that the various (application) factors may be weighed and integrated into a single numerical measure of performance. The cost measure of performance is a weighted probabilistic sum of cost over all error conditions. The probabilities involved are both the (application-dependent) probabilities of the various conditions and the (system-dependent) probabilities of error given these conditions. The cost of an error is assumed to be a function of the condition. So, in general, we have:

$$Cost = \sum_{\text{all Cnd}} \{ C_{\text{Error|Cnd}} \times P_{\text{Error|Cnd}} \times P_{\text{Cnd}} \}$$

where

$C_{\text{Error|Cnd}}$ = the cost of an error for condition = Cnd

$P_{\text{Error|Cnd}}$ = the (system) prob. of an error for condition = Cnd

P_{Cnd} = the (prior) probability of condition Cnd

There will be a single basic cost model for measuring speaker detection performance.

3.1 Speaker Detection Performance

3.1.1 The basic speaker detection cost model

The performance measure to be used for all speaker detection tests is the detection cost function, defined as a weighted sum of miss and false alarm error probabilities:

$$C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Miss|Target}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm|NonTarget}} \times (1-P_{\text{Target}})$$

The parameters of this cost function are the relative costs of detection errors, C_{Miss} and $C_{\text{FalseAlarm}}$, and the *a priori* probability of the specified target speaker, P_{Target}. The parameter values in Table 1 will be used as the primary evaluation of speaker recognition performance for all speaker detection tasks.

Table 1 Speaker Detection Cost Model Parameters for the primary evaluation decision strategy

C_{Miss}	$C_{\text{FalseAlarm}}$	P_{Target}
10	1	0.01

3.1.2 Normalization of speaker detection cost

One of the advantages of using a cost model is that it can be easily applied to different applications simply by changing the model parameters. On the other hand, a potential disadvantage of using cost as a performance measure is that it gives values that often lack intuitive meaning. To improve the intuitive value of the cost measure, we normalize the cost by dividing by C_{Default}, which is defined to be the best cost that could be obtained without processing the input data (i.e., by always making the same decision, namely either to accept or to reject the segment speaker as being the target speaker, whichever gives the lowest cost):

$$C_{\text{Default}} = \min\{ C_{\text{Miss}} \times P_{\text{Target}}, \quad C_{\text{FalseAlarm}} \times P_{\text{NonTarget}} \}$$

and

$$C_{\text{Norm}} = C_{\text{Det}} / C_{\text{Default}}$$

This default normalizing cost represents zero value, because this is the cost for a system that provides no information. The range of values for the normalized cost is:

$$C_{\text{Norm}} \in \{ 0, \ 1 + \max(C_{\text{Ratio}}, C_{\text{Ratio}}^{-1}) \}$$

where

$$C_{\text{Ratio}} = [C_{\text{Miss}} \times P_{\text{Target}}] / [C_{\text{FalseAlarm}} \times (1 - P_{\text{Target}})]$$

4 EVALUATION CONDITIONS

4.1 Evaluation of Speaker Detection

Speaker detection performance will be evaluated in terms of the detection cost function. The cost function will be computed over an ensemble of speech segments selected to represent a statistical sampling of conditions of interest. For each of these segments a set of speaker identities will be assigned as test hypotheses. Each of these hypotheses must be independently judged as "true" or "false", and the correctness of these decisions will be tallied.[6]

In addition to the actual detection decision, a decision score will also be required for each test hypothesis. This decision score will be used to produce detection error tradeoff curves, in order to see how misses may be traded off against false alarms.[7]

4.1.1 Cross-Condition Training and Test Data

The one-speaker detection and two-speaker detection conditions with limited data are different in both the training and the test data used, but draw their data from the same underlying switchboard cellular source. Therefore, to better understand performance characteristics and to help attribute differences in performance correctly to training versus test, there will be included some one-speaker test segment trials that use two-speaker models, and some two-speaker test segment trials that use one-speaker models.

[6] This means that an explicit speaker detection decision is required for each trial. Explicit decisions are required because the task of determining appropriate decision thresholds is a necessary part of any speaker detection system and is a challenging research problem in and of itself.

[7] Decision scores from the various target speakers will be pooled before plotting detection error tradeoff curves. Thus it is necessary to normalize scores across speakers to achieve satisfactory detection performance.

4.1.2 One-Speaker Detection – limited data

4.1.2.1 Training Data

Training data for each speaker will consist of about two minutes of speech from a single conversation. The actual duration of the training data used will vary slightly from this nominal value so that whole turns may be included whenever possible. Actual durations will, however, be constrained to lie within the range of 110-130 seconds.

4.1.2.2 Test Data

Each test segment will be extracted from a 1-minute excerpt of a single conversation and will be the concatenation of all speech from the subject speaker during the excerpt. The duration of the test segment will therefore vary, depending on how much the segment speaker spoke.

Evaluation trials will include both "same number" and "different number" tests. Evaluation trials may also include some cross-sex conversations, in which the model speaker and test speaker are of opposite sexes.

The primary evaluation conditions are:

1. "different number" tests *(unless there is an insufficient quantity of "different number" tests).*
2. The model is a one-speaker model.
3. The speech duration is between 15 and 45 seconds.
4. Both speakers are of the same sex.

Results will be tabulated separately for male and female model speakers. There will be no cross-sex tests.

4.1.3 Two-Speaker Detection

4.1.3.1 Training Data

Three whole conversations (minus some introductory comments) will be used for training. In contrast with the one-speaker training condition, however, the two sides of each conversation will be summed together, and both the model speaker and that speaker's conversation partner will be represented in this conversation. Thus the challenge is to separate the speech of the two speakers and then to decide (correctly) which is the model speaker. To make this challenge feasible, the training conversations will be chosen so that all speakers other than the model speaker are represented in only one conversation. Thus the model speaker, who is represented in all three conversations, is the only speaker to be represented in more than one.

4.1.3.2 Test Data

Each test segment will have a duration of nominally 60 seconds and will be the sum of the two sides of a conversation. The actual duration will vary from nominal, so that the test segment begins and ends on a speaker turn boundary. Actual test segment duration will, however, be constrained to lie within the range of 59-61 seconds. Note that the duty cycle of a segment speaker may vary from 0% to 100%.

There are three possible cases with respect to gender for each test segment: both speakers are male; both are female; or one is male and one female. Performance will be computed and evaluated separately for each of these three cases, but the system will not be given prior knowledge detailing the gender mix of the test segment. (Automatic gender detection may be used, of course.)

The primary evaluation conditions are:

1. "different number" tests *(unless there is an insufficient quantity of "different number" tests).*
2. The model is a two-speaker model.
3. The speech duration is 15-45 seconds for both speakers.
4. Both speakers are of the same sex.

4.1.4 One-Speaker Detection – extended data

This section outlines the conditions for the one-speaker detection task with extended training and test data. The entire SwitchBoard-II corpus phases 2 and 3 will be used for this evaluation. In addition several types of auxiliary data derived from the acoustic data will be made available to those who wish to use them. These will include

- automatically generated word transcriptions from a real-time recognizer (the same transcriptions as provided in 2002)
- automatically generated phone level transcriptions for five different languages (used by one system last year)
- pitch track estimates (from SRI)[8]
- base GMM-UBM scores (from MIT-Lincoln Lab)
- automatically generated handset type labels (from MIT-Lincoln Lab)

Researchers from several sites who have done previous work on this problem are generously supplying these information sources for the test corpora in support of a collaborative effort to achieve significant progress. Participants from other sites are invited to suggest other useful information sources which they might be able to supply for the benefit of the research community.

4.1.4.1 Training Data

Speaker training data will comprise all of one or more conversation sides for a given model speaker. A jackknife scheme that rotates training and test data will be used in order to provide an adequate number of tests. In order to provide unbiased results, models must exclude test conversation sides from target speakers and all data from impostor speakers. This information will be provided in index files that must be used to control the evaluation. Instructions are given in section 8.2.2 for the use of this index file information.

Various training options exist. The acoustic data may be used alone, the transcriptions (ASR) and other auxiliary data (in any combination) may be used alone, or all data may be used in combination. Note that the conversation sides and the transcriptions are presented in their entirety, without excision or deletion.

[8] See Kemal Sonmez, Elizabeth Shriberg, Larry Heck & Mitchel Weintraub (1998), Modeling Dynamic Prosodic Variation for Speaker Verification, Proc. ICSLP, vol. 7, pp. 3189-3192.

4.1.4.2 Test Data

The task is one-speaker detection. One whole conversation side will serve as the test segment. As in training, the acoustic data may be used alone, the transcriptions (ASR) and other auxiliary information (in any combination) may be used alone, or all information may be used in combination. And as in training, the data are presented in their entirety for the whole conversation side, without excision or deletion.

Results will be evaluated as a function of the amount target speaker training data, the handset types, and speaker sex. Some cross-sex trials will also be included.

5 DEVELOPMENT DATA

The evaluation data for the speaker detection conditions of last year's evaluation will serve as the development data for corresponding parts of this year's evaluation. Please refer to last year's evaluation plan for details.[9]

6 EVALUATION DATA

The LDC will provide a license agreement which participating sites must sign governing the use of the data for the speaker detection conditions with limited data. The license terms may depend upon whether or not a site is a current LDC member.

For the extended data condition, participating sites are expected to acquire the audio corpora used from the LDC.

6.1 One-speaker detection – limited data

The evaluation data will be drawn from the Switchboard Cellular Corpus, Part 2. All conversations will have been processed through echo canceling software before being used to create training and test segments.

Training and test segments will be constructed by concatenating consecutive turns of the desired speaker (but note section 4.1.1). Each such segment will be stored as an 8-bit mu-law continuous speech signal in a separate SPHERE file. The SPHERE header of each such file will contain some auxiliary information as well as the standard SPHERE header fields.

There will be about 400 target speakers and about 3500 test segments. Each test segment will be evaluated against 11 hypothesized speakers of the same sex as the segment speaker.

6.2 Two-speaker detection – limited data

The same data will be used as in one-speaker detection with limited data. The training and test data will have the two sides of the conversation summed together (but note section 4.1.1). The training segments will consist of whole conversation sides. The test segments will each be approximately one minute in duration. The one-speaker test segments (section 6.1) will be concatenated one-speaker excerpts from these segments.

There will be about 400 target speakers and about 1500 test segments. Each test segment will be evaluated against 22 hypothesized speakers.

[9] The year 2001 speaker recognition evaluation plan may be accessed from http://www.nist.gov/speech/tests/spk/2002/doc/

6.3 One-speaker detection – extended data

The Switchboard-II Corpus, Phases 2 and 3, will again serve as the evaluation data for the extended data evaluation. Speech recognition transcriptions and other auxiliary data derived from the audio will also be made available. (See section 4.1.4). The audio data must be obtained from the LDC, which makes it available for sale to non-members.[10]

7 EVALUATION RULES

In order to participate in the 2003 speaker recognition evaluation, a site must complete, in its entirety, at least one complete evaluation of one of the three evaluation task conditions.[11]

All participants must observe the following evaluation rules and restrictions:

- Each decision is to be based only upon the specified test segment and target speaker. Use of information about other test segments and/or other target speakers is **not** allowed.[12] For example:
  - Normalization over multiple test segments is **not** allowed.
  - Normalization over multiple target speakers is **not** allowed.
  - Use of evaluation data for impostor modeling is **not** allowed (except for the extended data test as indicated in the index files).
- The use of manually produced transcripts or other information for training is **not** allowed.
- Knowledge of the sex of the *target* speaker (implied by data set directory structure as indicated below) **is** allowed.
- Knowledge of the sex of the speaker(s) in the *test segment* is **not** allowed (except as determined by automatic means, of course).
- Listening to the evaluation data, or any other experimental interaction with the data, is **not** allowed before all test results have been submitted. This applies to training data as well as test segments.
- Knowledge of the "*start*" and "*ending*" times that were used to construct the test segments (found in the SPHERE header -- see SPHERE Header Information, below) **is** allowed.
- Knowledge of any information available in the SPHERE header **is** allowed.

8 EVALUATION DATA SET ORGANIZATION

8.1 One-speaker detection – limited data

The organization of the one-speaker limited data will be:

[10] Corpus information may be found on the LDC website: http://www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone

[11] Participants are encouraged to do as many tests as possible. However, it is absolutely imperative that results for all of the test segments and target speakers in a test be submitted in order for that test to be considered valid and for the results to be accepted.

[12] This means that the technology is viewed as being "application-ready". Thus a system must be able to perform speaker detection simply by being trained on a specific target speaker and then performing the detection task on whatever speech segment is presented, without the (artificial) knowledge of other test data.

- A single top level directory used as a unique label for the disk: "**sp03c1sN**" where N is a digit identifying the disc.
- Under which there will be three directories "**train**" "**test**" and "**doc**"
- "**train**" and "**test**" will both contain "**male**" and "**female**" subdirectories which in turn will contain the appropriate data
- Training data will be a SPHERE formatted file. The file name will be a four-digit speaker ID with a ".sph" extension.
- Test data will be pseudo random names consisting of four characters followed by a ".sph" extension.
- Each **test** subdirectory will contain an index file (detectN.ndx) that identifies the evaluation trials to be performed, where "N" is a digit. (This will provide different index names even if the male test data is spread out across more than 1 disc.)

8.2 Two-speaker detection – limited data

The evaluation data set organization of the cellular telephone data will be:

- A single top level directory used as a unique label for the disk: "**sp03c2sN**" where N is a digit identifying the disc.
- Under which there will be three directories "**train**" "**test**" and "**doc**"
- The "**train**" directory will contain all the speech data to be used for training the various models. There will also be two lists ("m_train.lst" and a "f_train.lst"). Each list will contain one record per line, with a record consisting of a model id followed by three sphere waveforms to be used to create the model. Each field in the training lists will be separated by white space.
- The "**test**" will contain the two-speaker evaluation test data and will be pseudo random names consisting of four characters followed by a ".sph" extension.
- The **test** subdirectory will contain an index file (detectN.ndx) that identifies the evaluation trials to be performed.

8.3 One-speaker detection – extended data

The SwitchBoard-II phases 2 and 3 corpora will serve as the primary data set for the extended data evaluation. In addition, NIST will provide a speaker-conversation table and an evaluation control file to support system development and to define the evaluation test. These two files are available via web access.[13]

8.3.1 The speaker-conversation table

The speaker-conversation table is a file that gives the conversation-side filenames for each speaker in the corpus.[14] The format for these records is:

speaker = SPKR-ID, sex = S, **conversation-sides** = {CNV-SIDE}

where:

SPKR-ID is the speaker identifier,

S is either **M** (for male) for **F** (for female), and

CNV-SIDE is a conversation side identifier. CNV-SIDE is defined to be the identification number of a conversation followed by either A (for the caller) or B (for the callee). For example, "1234A". {CNV-SIDE} is the set of all conversation sides in the SwitchBoard corpus for which speaker SPKR-ID is the speaker, whitespace separated.

8.3.2 The evaluation control file

A single evaluation control file will be used to supervise the evaluation. This file will control the creation of models and define the testing of those models. The structure of the control file will accommodate systems that create a background model in addition to the obligatory target model. This control of background model creation is necessary to ensure unbiased testing, because of the jackknifing of training and test data within the test corpus.

Because of the jackknifing of training and test data, multiple background models will need to be created during the course of the extended data test. Recognizing that background model creation can be the most time consuming part of system development, the evaluation control file will be structured to reduce the number of background models needed.

It is planned to use the same evaluation control file as in the 2002 evaluation. It is possible that some modifications or additions to the file may be adopted, however. Participants should obtain the current version of the file from the evaluation web site.

The evaluation control file will contain records of three different types. The first type will be the background model specification record. Then, for each background model specification there will be one or more target model specification records. Finally, for each target model there will be one or more trial specification records.

The format for the background model specification record is:

BM: excluded-speakers = {SPKR-ID}

where:

SPKR-ID is a speaker identifier, and {SPKR-ID} is the set of speakers that must be **excluded** from the background model, whitespace separated. (These speakers are those from whom test data will be drawn.)

The format for the target model specification record is:

TM: MODEL-ID **target-sides** = {CNV-SIDE}

where:

MODEL-ID is a unique model identifier. This is required for the extended data task because there are multiple models for each speaker in this task. Having a unique model ID is therefore needed in order to uniquely associate a particular detection output with the model that produced it.

{CNV-SIDE} is the set of conversation sides (spoken by the target speaker) from which the target model is to be created, whitespace separated. These conversation sides are the **only** data that may be used to create the target speaker model.

[13] At http://www.nist.gov/speech/tests/spk/2002/extended-data/

[14] It is mandatory to use the information in this table.

There will be no more than 30 of these conversation sides per model.

The format for the trial specification record is:

test-sides = {CNV-SIDE}

where:

{CNV-SIDE} is the set of conversation sides to be used as test segments, whitespace separated, with one trial per test segment. {CNV-SIDE} contains data for both the target speaker and impostors.

The evaluation control file will specify no more than 10 different background models, no more than a total of 5,000 different target models, and no more than an aggregate total of 60,000 different trials. Some cross-sex trials will be included in the evaluation.

9 FORMAT FOR SUBMISSION OF RESULTS

Results for each test must be stored in a single file, according to the formats defined in this section. The file name should be intuitively mnemonic and should be constructed as "SSS_N", where

- SSS identifies the site, and
- N identifies the system.

9.1 Speaker Detection Test Results

Sites participating in the one-speaker evaluation tests must report results for whole tests, including all of the test segments. These results must be provided to NIST in a single results file using a standard ASCII format, with one record for each decision. Each record must document its decision with the target identification, test segment identification, and decision information. Each record must contain seven fields[15], separated by white space and in the following order:

1. The sex of the target speaker – **M** or **F**
2. The target model ID[16] (*a four digit speaker number plus, for the extended data condition, the model size*)
3. The test – (**1L** for one-speaker detection – limited data, **2L** for two-speaker detection – limited data, **1E** for one-speaker detection – extended data.)
4. The test segment identifier. This is the test segment file name (*excluding directory and file type*) for all conditions except the extended data condition, in which case it is the conversation-side ID.
5. The decision – **T** or **F** (*is the target speaker judged to be the same as the speaker in the test segment*)
6. The score (*where the more positive the score, the more likely the target speaker*)

10 SYSTEM DESCRIPTION

A brief description of the system(s) (the algorithms) used to produce the results must be submitted along with the results, for each system evaluated. It is permissible for a single site to submit multiple systems for evaluation for a particular test. In this case, however, the submitting site must identify one system as the "primary" system for the test prior to performing the evaluation.

Sites must report the CPU execution time that was required to process the test data, as if the test were run on a single CPU. Sites must also describe the CPU and the amount of memory used.

11 SCHEDULE

The deadline for signing up to participate in the evaluation is April 1, 2003.

The evaluation data set CD-ROM's will be distributed by NIST on April 14, 2003.

The deadline for submission of evaluation results to NIST is May 12, 2003.

Room reservations for the follow-up workshop must be received by (a date to be determined).

The follow-up workshop will be held on June 24-25, 2003 at the University of Maryland University College in College Park, Maryland. Those participating in the evaluation are expected to present and discuss their findings at the workshop.

12 GLOSSARY

Trial – The individual evaluation unit for each task involving a test segment and (except for segmentation) a hypothesized speaker.

Target (true speaker) trial – A trial in which the actual speaker of the test segment *is in fact* the target (hypothesized) speaker of the test segment.

Non-target (impostor) trial – A trial in which the actual speaker of the test segment *is in fact not* the target (hypothesized) speaker of the test segment.

Target (model) speaker – The hypothesized speaker of a test segment, one for whom a model has been created from training data.

Non-target (impostor) speaker – A hypothesized speaker of a test segment who is in fact not the actual speaker.

Segment speaker – The actual speaker in a test segment.

Turn – The interval during a conversation during when one participant speaks while the other remains silent.

[15] The seventh field is optional except for where a no-decision option is included.

[16] The target model ID is simply the speaker ID, except for the extended data task. For the extended data task, detection trials are performed for multiple models for each speaker, and therefore a target model ID, as specified in section 8.3.1, is required to uniquely identify the trials.