The 1996 “Hub-4” Annotation Specification for Evaluation of Speech Recognition on Broadcast News

Introduction

The annotation conventions provided herein are intended to support the Hub-4 R&D effort and evaluation. The purpose of the annotation is to associate certain signal, speaker, and content conditions with the speech and its transcription. This is done using SGML markup tags. These tags identify the speech using start/end time attributes, and they identify the transcription by enclosing it within the span of the tags. SGML markup conventions are used to maximize portability and to exploit existing knowledge and software tools.

Annotation Objectives

The Hub-4 Evaluation Specification requires that speech recognition performance be measured along several dimensions involving speaker characteristics, speech content, and the quality of the recording environment and transmission medium. For the Unpartitioned Evaluation (UE) the annotation must merely delineate those sections of a recording for which the speech is to be recognized. For the Partitioned Evaluation (PE), the annotation must in addition identify those factors relevant to speech recognition and to the evaluation of recognition performance. There are six such factors. They are:

Episode – An Episode serves to identify the recording of a particular broadcast of a program on a certain date and at a certain time.

Section – A Section serves to identify a particular *portion* and/or *story* within an episode. Sections divide an episode into untranscribed portions (such as commercials and sports results) and into distinct stories which are identified by specific events and/or topics.

Speaker – A Speaker’s name serves to identify each speaker uniquely. This name may be used as a key to look up Speaker attributes in an associated SGML-encoded table of Speakers.

Mode – The Mode of speaking may assume one of two values, either “Spontaneous” or “Planned” (as opposed to “Read”). The Hub-4 evaluation specification alludes to a contrast between “spontaneous” speech and “read” speech. The term “read”, however, does not seem to be a good division for broadcast speech, and whether speech is truly “read’ is often difficult to determine. For this reason, the notion of “Planned” speech is substituted. The division between “Spontaneous” and “Planned” is in terms of the characteristics that distinguish spontaneous from read speech, namely whether the speech exhibits disfluencies, such as restarts and hesitations (filled and/or unfilled). Planned speech is defined to be largely free of such phenomena (as is typical of read speech), whereas spontaneous speech is rife with them.

Fidelity – Fidelity serves to identify the quality of the recording environment and transmission medium. Fidelity may assume one of three values, namely “High”, “Medium”, or “Low”. High corresponds to the kind of speech to be expected in a broadcast studio: high bandwidth and low distortion. Low corresponds to the kind of speech to be expected over the telephone: low bandwidth and/or distorted. Medium corresponds to speech that cannot be fairly classified as either High or Low.

Background – Background serves to identify the nature of (undesired) acoustical input that is mixed with the (desired) speech signal. Background may assume one of three values, namely “Speech”, “Music” or “Other”. “Speech” indicates the presence of significant background speech; “Music” indicates the presence of significant background music; and “Other” indicates the presence of a significant background signal other than speech or music. (The lack of a background designation indicates that there is no significant background signal.)

The objective of the Hub-4 annotation is to identify these six factors and their associated times, and to provide this information, along with the transcription, in an accessible form to the researchers and their systems. This information, provided by LDC according to the specifications defined here, comprises two parts, namely the primary annotation file and an auxiliary speaker information file. In addition, NIST will provide a software tool to transform this primary annotation file into a form that maps time intervals to the specific focus conditions as defined in the Hub-4 evaluation specification.

Basic temporal organization – the “Segment” and the “Partition”

The Hub-4 Evaluation Specification defines “segment” to be a continuous time span of audio signal during which the evaluation conditions remain constant. Thus “Segment” will serve as the basic SGML tag which will be used to organize the transcription. Topic, Speaker, Fidelity and Mode are the evaluation conditions that will determine Segment boundaries. A Segment boundary occurs whenever one of these evaluation conditions changes.

There is one other evaluation condition, namely Background, which is not used in determining segment boundaries. There are several reasons for this: The background is logically independent of the speech signal; changes in background are often asynchronous with the other conditions; and it is more efficient to annotate the background separately. Thus, LDC will identify and annotate Segment boundaries and Background conditions independently of each other, and changes in the Background will be ignored in determining Segment boundaries. NIST will provide a partition mapper tool, however, that will automatically transform the LDC annotation files and produce “Partitions”, based on Segment and Background annotation, during which the evaluation conditions are indeed constant. (Please refer to the description of this tool later in this document.) Note that the Partition boundaries output by NIST’s partition mapper (which will be used for evaluation) will be different from the Segment boundaries in the LDC annotation, and the number of Partitions will generally be larger than the number of Segments.

The SGML tags and their attributes

Episode – <**Episode**> is a spanning tag, terminated by </**Episode**>. It spans all of the annotation and transcription information associated with a particular episode, and it may contain <**Section**>, <**Background**> and <**Comment**> tags within its span. The attributes associated with each Episode are:

Filename: The name of the file containing the episode’s audio signal.

Scribe: The name of the transcriber who produced the annotation and the transcription.

Program: The name of the program that produced the episode. (E.g., “NPR_Marketplace”)

Date: The date and time of the episode broadcast, in “YYMMDD:HHMM” format. (E.g., “960815:1300”.)

Version: The version number of the annotation of this episode, starting with “1”. Each time the annotation is revised, the version number is incremented by 1.

Version_Date: The (last) date and time of annotation/transcription input to this annotation.

Section – <**Section**> is a spanning tag, terminated by </**Section**>. It spans all of the annotation and transcription information associated with a particular section of an episode, and it may contain <**Segment**>, <**Background**> and <**Comment**> tags within its span. It must be contained within the span of an <**Episode**>. The attributes associated with each Section are:

S_time: The start time of the Section, measured from the beginning of the Episode in seconds.

E_time: The end time of the Section, measured from the beginning of the Episode in seconds.

Type: One of the labels "Story", "Filler", "Commercial", "Weather_Report", "Traffic_Report", "Sports_Report", or "Local_News". For the current Hub-4 effort, Commercials and Sports_Reports will not be transcribed and will therefore contain no Segments. Sections of all other Types will be transcribed and will be included in the evaluation.

Topic: An identification of the event or topic discussed in the Section. For example, "TWA flight 800 disaster". Topic is optional and is not currently being supplied by LDC. (Future use and value of Topic will require additional guidance on how to define it.)

Segment – <**Segment**> is a spanning tag, terminated by </**Segment**>. It spans all of the annotation and transcription associated with a particular Segment, and it may contain <**Sync**>, <**Background**> and <**Comment**> tags within its span, as well as the transcription text. The <**Segment**> tag must be contained within the span of a <**Section**>. (Segment information is allowable only for the PE.) The attributes associated with each Segment are:

S_time: The start time of the Segment, measured from the beginning of the Episode in seconds.

E_time: The end time of the Segment, measured from the beginning of the Episode in seconds.

Speaker: The speaker's name.

Mode: One of the labels "Spontaneous" or "Planned".

Fidelity: One of the labels "High", "Medium" or "Low.

Sync – <**Sync**> is a non-spanning tag that provides transcription timing information within a Segment. It is positioned within the transcription and gives the time at that point. The <**Sync**> tag must be contained within the span of a <**Segment**>. Sync is a side-effect of the transcription process and is being provided for potential convenience. Sync has a single attribute, namely Time:

Time: The time at this point in the transcript, measured from the beginning of the Episode in seconds.

Comment – <**Comment**> is a spanning tag, terminated by </**Comment**>. It spans a free-form text comment by the transcriber, but no other SGML tags. The <**Comment**> tag must be contained within the span of an <**Episode**>.

Background – <**Background**> is a non-spanning tag that provides information about a particular (single) background signal, specifically regarding the onset and offset of the signal. This information is synchronized with the transcript by positioning the Background tag at the appropriate point in the transcription. (<**Background**> tag locations and times will be positioned at word boundaries so that the word within which the background noise starts or ends will be included in the span of the background noise.) The <**Background**> tag must be contained within the span of an <**Episode**>. The attributes associated with each Background tag are:

Time: The time at this point in the transcript, measured from the beginning of the Episode in seconds.

Type: One of the labels "Speech", "Music" or "Other".

Level: One of the labels "High", "Low" or "Off". This attribute indicates the level of the background signal *after* Time. Thus High or Low implies that the signal *starts* at Time, while Off implies that the signal *ends* at Time.

Foreign (non-English) speech will be labeled as background speech and not transcribed, even if it appears to be in the foreground. The exception to this is that occurrences of borrowed foreign words or phrases, when used within English speech, are transcribed.

Overlap – <Overlap> is a spanning tag, terminated by </**Overlap**>. It is used to indicate the presence of simultaneous speech from another foreground speaker.[1] This information is synchronized with the transcript by positioning the Overlap tag at the appropriate point in the transcription. (<**Overlap**> tag locations and times will be positioned at word boundaries so that the word within which the overlap starts or ends will be included in the span of the overlap.) The <**Overlap**> tag must be contained within the span of a <**Segment**>. The attributes associated with each Overlap tag are:

S_time: The start time of the Overlap, measured from the beginning of the Episode in seconds.

E_time: The end time of the Overlap, measured from the beginning of the Episode in seconds.

For example:

Speaker A: ... It was a tough game <Overlap S_time=101.222 E_time=102.111> # but very exciting # </Overlap>

Speaker B: <Overlap S_time=101.230 E_time=102.309> # Yes it was # </Overlap>

In this example, Speaker B broke into Speaker A's turn. Note that the Overlap times don't coincide exactly because they have been time-aligned to the most inclusive word boundaries for each speaker Segment involved in the overlap.

Expand – <Expand> is a spanning tag, terminated by </**Expand**>. It is used to indicate the expansion of a transcribed representation, such as a contraction, to a full representation of the intended words that underlie the transcription. The **<Expand>** tag spans the word(s) to be expanded and must be contained within the span of a **<Segment>**. Expand has a single attribute, namely E_form:

E_form: The expanded form of that portion of the transcription spanned by the Expand tag.

To illustrate, here is a simple example: **I <Expand E_form="do not"> don't </Expand> think <Expand E_form="he is"> he's </Expand> lying.** Note that the transcribed words remain unchanged, while the attribute E_form indicates the correct expansion of the spanned words. Note also that E_form resolves potential ambiguity (such as whether "he's" should be expanded to "he is" or "he has").

Noscore - <Noscore> is a spanning tag, terminated by **</Noscore>**. It is used to explicitly exclude a portion of a transcription from scoring. The **<Noscore>** tag spans the word(s) to be excluded and must be contained within the span of a **<Segment>**. Noscore has 3 attributes:

Reason: Short free-form text string containing an explanation of why the tagged text has been excluded from scoring. The string must be bounded by double quotes.

S_time: The start time of the excluded portion, measured from the beginning of the Episode in seconds.

E_time: The end time of the excluded portion, measured from the beginning of the Episode in seconds.

For example:

<Noscore Reason="Mismatch between evaluation index and final transcript" S_time=1710.93 E_time=1711.71> ... text to be excluded ... </Noscore>

[1] **Overlap** is an improved SGML notation that supersedes the use of "#" to indicate the start and end of regions of overlap. For backward compatibility, the "#" characters will be retained for now, even though they are redundant with, and contain less information than, the **<Overlap>** tags.

The Transcription

Character set and line formatting

The transcription text will consist of mixed-case ASCII characters. Only alphabetic characters and punctuation marks will be used, along with the bracketing characters listed below. Line breaks may be inserted within the text, to keep lines less than 80 characters wide and to separate the transcription text from SGML tags. (Transcription text will not be entered on the same line with any SGML tag.)

Numbers, Acronyms and Abbreviations

Numbers are transcribed as words (e.g. “ninety six” rather than “96”). Acronyms are transcribed as upper-case words when said as words (e.g., “AIDS”). When said as a sequence of letters, acronyms are transcribed as a sequence of space-separated upper-case letters with periods (e.g., “C. N. N.”). Except for “Mr.” and “Mrs.”, abbreviations are not used. However, words that are spoken as abbreviated (e.g., “corp.” rather than “corporation”) are spelled that way.

Special Bracketing Conventions

Single word tokens may be bracketed to indicate particular conditions as follows:

… indicates a neologism – the speaker has coined a term. E.g., **Mediscare**.

+…+ indicates a mispronunciation. The intended word is transcribed, regardless of its pronunciation. E.g., +ask+ rather than +aks+. (Variant pronunciations that are intended by the speaker, such as “probly” for the word probably, are not bracketed.)

[…] indicates (a one-word description of) a momentary intrusive acoustic event not made by the speaker. E.g., [gunshot].

{…} indicates (a one-word description of) a non-speech sound made by the speaker. E.g., {breath}.

Sequences of one or more word tokens may be bracketed to indicate particular conditions as follows:

((…)) indicates unclear speech, where what was said isn’t clear. The parentheses may be empty or may include a best guess as to what was said.

… # indicates simultaneous speech. This occurs during interactions when the speech of two people who are being transcribed overlap. The words in both segments that are affected are bounded by # marks.

Other notations

@ indicates unsure spelling of a proper name. The transcriber makes a best guess and prefixes the name with the @ sign. E.g., Peter @Sprogus.

- indicates a word fragment. The transcriber truncates the word at the appropriate place an appends the - sign. E.g., bac-.

Punctuation

With the exception of periods (“.”), normal punctuation is permitted, but not required. Periods are used only after spelled out letters (N. I. S. T.) and in the accepted CSR abbreviations (Mr., Mrs., Ms.). They may not be used to end sentences. Instead, sentences may be delimited with semicolons.

Non-English speech

Speech in a foreign language will not be transcribed. This speech will be indicated using the "((…))" notation. However, for foreign words and phrases that are generally understood and in common usage (such as "adios"), these words will be transcribed with customary English spelling and will be treated as English.

The Annotation format

With the exception of <**Comment**>, the beginning mark of each spanning tag will be presented alone and complete on one line. The corresponding ending mark will also appear alone on a subsequent line. The <**Comment**> units are often brief, but they are free to extend to multiple lines. Within the beginning marks of spanning tags, all attribute value assignments will be bounded by spaces (except the last, where a space isn't needed before the closing ">"). Attributes containing spaces or other non-alphanumeric characters must be enclosed in quotes. Here is an example of annotation:

```
<Episode Filename=f960531.sph Scribe=Stephanie_Kudrac Program=CNN_Headline_News
Date="960531:1300" Version=1 Version_Date="960731:1730">

<Section S_time=0.28 E_time=105.32 Type=Commercial>

</Section>

<Background Time=111.27 Type=Music Level=High>

<Section S_time=116.55 E_time=124.92 Type=Filler Topic="lead-in">

<Segment S_time=117.61 E_time=121.06 Speaker=Announcer_01 Mode=Planned Fidelity=High>

Live from Atlanta with Judy Forton

</Segment>

<Segment S_time=121.95 E_time=124.92 Speaker=Judy_Forton Mode=Spontaneous Fidelity=High>

Lynn Vaughn is off today;  Thanks for joining us;

</Segment>

</Section>

<Section S_time=124.92 E_time=299.79 Type=Story Topic="U.S. - Israeli politics">

<Segment S_time=124.92 E_time=139.20 Speaker=Judy_Forton Mode=Planned Fidelity=High>

President Clinton has congratulated Israel's next

<Sync Time=127.74>

leader

<Background Time=128.30 Type=Music Level=Off>

and has invited him to the White House to talk about Middle East

<Sync Time=131.03>

peace strategies {breath} President Clinton called Benjamin Netenyahu just minutes

<Sync Time=135.04>

after he was declared the winner over Prime Minister Shimon Peres  {breath} Fred Saddler reports

</Segment>

<Background Time=139.65 Type=Other Level=Low>
```

<Comment> background noise and people </Comment>

<Segment S_time=141.32 E_time=154.88 Speaker=Fred_Saddler Mode=Planned Fidelity=Medium>

Never doubting that he would win, Benjamin Netenyahu came out on top

</Segment>

</Section>

</Episode>

Speaker Information

A list of speakers and their attributes will be stored in a separate SGML-structured file. This list will provide information for all of the speakers in the Hub-4 corpus. Two SGML tags are used for this purpose. Here is a definition of them and their associated attributes:

Speaker_list – <**Speaker_list**> is a spanning tag, terminated by </**Speaker_list**>. It spans all of the speaker information associated with a particular corpus, and it contains <**Speaker**> tags within its span. There is one attribute associated with the Speaker_list, which is:

Corpus_ID: The name/version of the corpus for which this Speaker_list applies.

Speaker – <**Speaker**> is a non-spanning tag which provides all of the information about a particular speaker through attribute values. The <**Speaker**> tag must be contained within the span of a <**Speaker_list**>. The attributes associated with each Speaker are:

Name: The speaker's name. (The speaker's name must be globally unique, within the subject corpus. If for example there are two John Smiths, then they might be represented in the Speaker_list as "John_Smith_1" and "John_Smith_2".) The speaker's name is the same as that referenced by a Segment's Speaker attribute. Thus each Segment's Speaker value may be used as a key into the Speaker_list to access information for that Speaker.

Sex: One of the values "Male" or "Female". This attribute may not be provided for some speakers, for example for children in cases where the sex is unknown and is uncertain based on listening.

Dialect: One of the labels "Native" or "Nonnative". The value, "Native", indicates a native speaker of North American English. The value, "Nonnative", indicates all other dialects. Future corpora may include finer distinctions such as "General_American", "Southern_American", "Black", "Hispanic", "British", etc.

Age: One of the labels "Juvenile", "Adult", or "Elderly".

Role: The role of the speaker. For example "Announcer", "Reporter", "Correspondent", "Fireman", "Police chief", "Airline attendant" or "Citizen". This attribute is optional and is provided at the discretion of the transcriber.

Here is an example of a Speaker_list:

<**Speaker_list** Corpus_ID=Hub4_96>

<**Speaker** Name=Announcer_01 Sex=Male Dialect=Native Age=Adult Role=Announcer>

<**Speaker** Name=Judy_Forton Sex=Female Dialect=Native Age=Adult Role=Reporter>

<**Speaker** Name=Fred_Saddler Sex=Male Dialect=Native Age=Adult Role=Correspondent>

</**Speaker_list**>

Focus Conditions for the 1996 CSR Hub-4 Evaluation

The following focus conditions determine the partitioning in the 1996 CSR Hub-4 Evaluation:

Condition	Dialect	Mode	Fidelity	Background
F0 – Baseline Broadcast:	native	Planned	High	Clean
F1 – Spontaneous Speech:	native	**Spontaneous**	High	Clean
F2 – Telephone Channels:	native	(any Mode)	**Medium/Low**	Clean
F3 – Background Music:	native	(any Mode)	High	**Music**
F4 – Degraded Acoustics:	native	(any Mode)	High	**Speech/Other**
F5 – Nonnative Speakers:	**nonnative**	Planned	High	Clean
FX – All Other Combinations:	–	–	–	–

Derivative Files for Evaluation and Research

NIST will provide a PERL tool (BN_filter) to parse the SGML annotation files and produce the following derivative files:

1. An SGML partitioned annotation (SPA) file which is re-tagged by evaluation focus condition (as described in the previous section) for research and diagnostics. (-f spa)

2. An STM file for evaluation results scoring. (-f stm)

3. Evaluation maps (index files) for implementing a PE or UE evaluation. (-f pem or -f uem)

Various other flags are required to indicate start time, end time, and other selectable parameters. See the documentation accompanying BN_filter for its use.

1. SGML Partitioned Annotation Format (SPA):

The -f spa option to BN_filter produces a focus-condition partitioned version of the original segment-based annotation file. This SGML Partitioned Annotation (SPA) file is based on a similar but slightly different set of tags as the original annotation. Episode and Section tags retain their meaning. Each Segment tag is transformed into one or more Partition tags depending on whether background condition changes occurred during the segment. The Partition tag has essentially the same meaning as the Segment tag, except that the Partition tag requires that a single focus condition prevail throughout its span. The focus condition is noted in the Partition tag via a Condition attribute. The Comment tags are discarded.

Here is the definition of the Partition tag:

Partition – <**Partition**> is a spanning tag, terminated by </**Partition**>. It spans all of the transcription associated with a particular Partition, and it contains within its span only the transcription text. The <**Partition**> tag must be contained within the span of a <**Section**>. The attributes associated with each Partition tag are:

S_time: The start time of the Segment, measured from the beginning of the Episode in seconds.

E_time: The end time of the Segment, measured from the beginning of the Episode in seconds.

Speaker: The speaker's name.

Mode: One of the labels "Spontaneous" or "Planned".

Fidelity: One of the labels “High”, “Medium” or “Low.

Speaker_Dialect: One of the labels “Native” or Nonnative”.

Condition: One of the labels “F0”, “F1”, “F2”, “F3”, “F4”, “F5” or “FX”.

Here is an example of the SPA output of BN_filter based on the previous annotation example:

```
<Episode Filename=f960531.sph Scribe=Stephanie_Kudrac Program=CNN_Headline_News
Date="960531:1300" Version=1 Version_Date="960731:1730">

<Section S_time=0.28 E_time=105.32 Type=Commercial>

</Section>

<Background Time=111.27 Type=Music Level=High>

<Section S_time=116.55 E_time=124.92 Type=Filler Topic="lead-in">

<Partition S_time=117.61 E_time=121.06 Speaker=Announcer_01 Mode=Planned Fidelity=High
Speaker_Dialect=Native Condition=F3>

LIVE FROM ATLANTA WITH JUDY FORTON

</Partition>

<Partition S_time=121.95 E_time=124.92 Speaker=Judy_Forton Mode=Spontaneous Fidelity=High
Speaker_Dialect=Native Condition=F3>

LYNN VAUGHN IS OFF TODAY THANKS FOR JOINING US

</Partition>

</Section>

<Section S_time=124.92 E_time=299.79 Type=Story Topic="U.S. - Israeli politics">

<Partition S_time=124.92 E_time=128.30 Speaker=Judy_Forton Mode=Planned Fidelity=High
Speaker_Dialect=Native Condition=F3>

PRESIDENT CLINTON HAS CONGRATULATED ISRAEL'S NEXT

<Sync Time=127.74>

LEADER

</Partition>

<Background Time=128.30 Type=Music Level=Off>

<Partition S_time=128.30 E_time=139.20 Speaker=Judy_Forton Mode=Planned Fidelity=High
Speaker_Dialect=Native Condition=F0>

AND HAS INVITED HIM TO THE WHITE HOUSE TO TALK ABOUT MIDDLE EAST

<Sync Time=131.03>

PEACE STRATEGIES PRESIDENT CLINTON CALLED BENJAMIN NETENYAHU JUST MINUTES

<Sync Time=135.04>

AFTER HE WAS DECLARED THE WINNER OVER PRIME MINISTER SHIMON PERES FRED
SADDLER REPORTS

</Partition>

<Background Time=139.65 Type=Other Level=Low>

<Background Time=139.65 Type=Music Level=Low>
```

```
<Partition S_time=141.32 E_time=154.88 Speaker=Fred_Saddler Mode=Planned Fidelity=Medium
Speaker_Dialect=Native Condition=FX>

NEVER DOUBTING THAT HE WOULD WIN BENJAMIN NETENYAHU CAME OUT ON TOP

</Partition>

</Section>

</Episode>
```

2. Segment Time Marked Format (STM):

The -f stm flag to BN_filter produces the format of the reference transcription used by the NIST Sclite scoring package. An STM file consists of a set of newline-separated records, one for each Partition. Each record provides annotation and transcription information for the Partition according to the following BNF format:

STM :== <F> <C> <S> <BT> <ET> [<LABEL>] <TRANSCRIPT>

Where:

<F> -> The waveform filename. Basename only – any path and/or extension is removed.

<C> -> The waveform channel (always 1 for Hub-4).

<S> -> The speaker id. Space-delimited, can consist of any other characters.

<BT> -> The begin time (in seconds) of the STM Partition.

<ET> -> The end time (in seconds) of the STM Partition.

<LABEL> -> A comma-separated list of subset identifiers enclosed in angle brackets. For example, “<O,F0>”. The PE condition is identified here in the second subfield. The other subfield identifies information used in scoring. See the manual page provided in the sclite distribution for further documentation.

<TRANSCRIPT> -> A whitespace-separated string of words which have been converted to SNOR format.

Records which begin with “;;” are comment records only. Here is an example of STM output from BN_filter for the above annotation example:

```
;; STM for File f960531.txt, Show CNN_Headline_News, Episode 960531:1300, Version 1 - 960731:1730

;;    Generated by BN_filter version 1.4

;;

;; Field 1: File ID

;; Field 2: Channel

;; Field 3: Speaker ID

;; Field 4: Start Time

;; Field 5: End Time

;; Field 6: Categories

;; CATEGORY "0" "" ""

;; LABEL "O" "Overall" "Overall"

;;
```

;; CATEGORY "1" "1996 Hub4 Focus Conditions" ""

;; LABEL "F0" "Baseline//Broadcast//Speech" ""

;; LABEL "F1" "Spontaneous//Broadcast//Speech" ""

;; LABEL "F2" "Speech Over//Telephone//Channels" ""

;; LABEL "F3" "Speech in the//Presence of//Background Music" ""

;; LABEL "F4" "Speech Under//Degraded//Acoustic Conditions" ""

;; LABEL "F5" "Speech from//Non-Native//Speakers" ""

;; LABEL "FX" "All other speech" ""

;; Field 7: SNOR Transcript

;;

f960531 1 Announcer_01 117.61 121.06 <O,F3> LIVE FROM ATLANTA WITH JUDY FORTON

f960531 1 Judy_Forton 121.95 124.92 <O,F3> LYNN VAUGHN IS OFF TODAY THANKS FOR JOINING US

f960531 1 Judy_Forton 124.92 128.30 <O,F3> PRESIDENT CLINTON HAS CONGRATULATED ISRAEL'S NEXT LEADER

f960531 1 Judy_Forton 128.30 139.20 <O,F0> AND HAS INVITED HIM TO THE WHITE HOUSE TO TALK ABOUT MIDDLE EAST PEACE STRATEGIES PRESIDENT CLINTON CALLED BENJAMIN NETENYAHU JUST MINUTES AFTER HE WAS DECLARED THE WINNER OVER PRIME MINISTER SHIMON PERES FRED SADDLER REPORTS

f960531 1 Fred_Saddler 141.32 154.88 <O,FX> NEVER DOUBTING THAT HE WOULD WIN BENJAMIN NETENYAHU CAME OUT ON TOP

3. Evaluation Maps (PEM and UEM):

The evaluation maps provide an index into the waveform files for implementing benchmark tests and include pertinent side information (if applicable). Sites will receive one such index for the Partitioned Evaluation (PEM) and another for the Unpartitioned Evaluation (UEM). Both indexes contain waveform-excerpt evaluation records. A PEM file contains pointers to excerpts in the waveforms to be evaluated along with the focus condition of the excerpts and a boolean flag indicating the beginning of a new Section. Each of these Partition records is followed by a newline and then a list of the factors yielding the focus condition in parenthesis. The factor list is then followed by a blank line.

For a PEM, these excerpts are identical to the Partitions as defined above for SPA files. A UEM file contains only start and end pointers for excerpts of the waveforms to be evaluated. Excerpts in the UEM are only partitioned to exclude untestable sections such as commercials.

PEM example

The following is an example PEM file corresponding to the example annotation:

;; PEM for File f960531.txt, Show CNN_Headline_News, Episode 960531:1300, Version 1 - 960731:1730

;;   Generated by BN_filter version 1.4

;;

;; Field 1: File ID

;; Field 2: Channel

```
;; Field 3: Speaker ID
;; Field 4: Start Time
;; Field 5: End Time
;; Field 6: Categories
;; CATEGORY "0" "1996 Hub4 Focus Conditions" ""
;; LABEL "F0" "Baseline//Broadcast//Speech" ""
;; LABEL "F1" "Spontaneous//Broadcast//Speech" ""
;; LABEL "F2" "Speech Over//Telephone//Channels" ""
;; LABEL "F3" "Speech in the//Presence of//Background Music" ""
;; LABEL "F4" "Speech Under//Degraded//Acoustic Conditions" ""
;; LABEL "F5" "Speech from//Non-Native//Speakers" ""
;; LABEL "FX" "All other speech" ""
;; Field 7: New Story (1=yes, 0=no)
;; Field 8: Condition Tags.
;;    The format of the condition tag is as follows:
;;    <COND_TAG> :== (Dialect=<DIALECT>,Mode=<MODE>,Fidelity=<FIDELITY>,Background_Music=<LEVEL>,Background_Bgspkr=<LEVEL>,Background_Other=<LEVEL>)
;;       where:
;;         <DIALECT>    :==  Native|Nonnative
;;         <MODE>       :==  Planned|Spontaneous
;;         <FIDELITY>   :==  High|Medium|Low
;;         <LEVEL>      :==  High|Low|Off
f960531 1 unknown_speaker 117.61 121.06 <F3> 1 (Dialect=Native,Mode=Planned,Fidelity=High,Background_Music=High,Background_Bgspkr=Off,Background_Other=Off)
f960531 1 unknown_speaker 121.95 124.92 <F3> 0 (Dialect=Native,Mode=Spontaneous,Fidelity=High,Background_Music=High,Background_Bgspkr=Off,Background_Other=Off)
f960531 1 unknown_speaker 124.92 128.30 <F3> 1 (Dialect=Native,Mode=Planned,Fidelity=High,Background_Music=High,Background_Bgspkr=Off,Background_Other=Off)
f960531 1 unknown_speaker 128.30 139.20 <F0> 0 (Dialect=Native,Mode=Planned,Fidelity=High,Background_Music=Off,Background_Bgspkr=Off,Background_Other=Off)
f960531 1 unknown_speaker 141.32 154.88 <FX> 0 (Dialect=Native,Mode=Planned,Fidelity=Medium,Background_Music=Low,Background_Bgspkr=Off,Background_Other=Low)
```

UEM file example

The following is an example UEM file corresponding to the example annotation:

```
;; UEM for File f960531.txt, Show CNN_Headline_News, Episode 960531:1300, Version 1 - 960731:1730
;;     Generated by BN_filter version 1.4
;;
;; Field 1: File ID
;; Field 2: Channel
;; Field 3: Speaker ID
;; Field 4: Start Time
;; Field 5: End Time
f960531.txt 1 116.55 299.79
```

Restrictions on the release of evaluation maps

The complete annotations for the evaluation data will not be made available to participants in the evaluation until immediately after the test results due date. NIST will provide only the above-described PEM and UEM evaluation maps along with the evaluation waveforms. After the sites complete the evaluation, the annotation files for the evaluation data will be supplied, to support diagnostic study of the results.