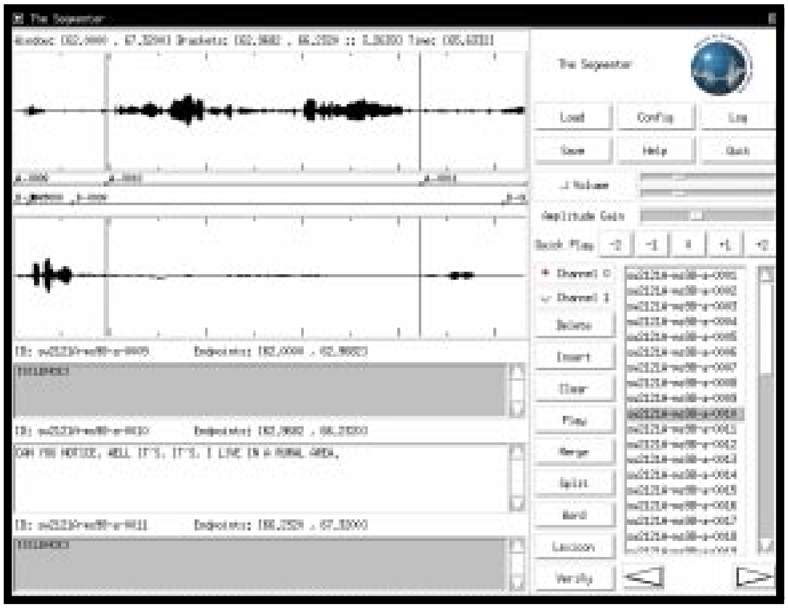
INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

# Rules and Guidelines for Transcription and Segmentation of the SWITCHBOARD Large Vocabulary Conversational Speech Recognition Corpus

Version 7.1

October 1, 1998



[SILENCE]

CAN YOU NOTICE, WELL IT'S, IT'S, I LIVE IN A RURAL AREA.

[SILENCE]

by,

J. Hamaker, Graduate Research Assistant
Y. Zeng, Graduate Research Assistant
J. Picone, Ph.D., Associate Professor

**Institute for Signal and Information Processing**
Institute for Signal and Information Processing
Mississippi State University
Box 9571, 216 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-3149, Fax: 601-325-3149
email: {hamaker, zeng, picone}@isip.msstate.edu

# General Instructions for SWITCHBOARD Transcriptions

This document is structured into two sections: the original SWITCHBOARD (SWB) transcription guidelines and the ISIP modifications to this standard. Historically, the problem with any SWB convention document has been that the data delivered does not conform to the guidelines. Hence, the ISIP modifications are somewhat a documentation of what conventions are embedded in the current corpus, along with some new conventions based on the goals of our project. The ISIP modifications appear first followed by the SWB standard. If a particular issue is not covered in the ISIP amendments section, then assume we are following the original SWB convention.

## 1. Appended Instructions

The following guidelines for segmentation and transcription of SWB take precedence over the original SWB transcription conventions supplied by LDC (and described in Section 2).

### 1.1. Segmentation

The original goal of this project was to provide a new segmentation of the database to support improved acoustic training for speech recognition. It is important to remember this goal when discussing the challenging problem of SWB segmentation. Note that we do not pay attention to turns and such linguistic phenomena in performing the segmentation. Our segmentation will be largely based on the acoustic data.

Conversations will be broken into a sequence of segments which we refer to as utterances. Utterances will consist of either speech padded by 0.5 secs of silence on each side, or consist of only silence (background noise). Further, a design goal of the project is that an utterance be no more than 15 seconds in length. Ideally, breakpoints will be inserted at natural linguistic points in the utterance such as sentence or phrase boundaries. When no suitable boundary can be found, we progressively relax the requirement that the silence padding be 0.5 seconds in duration. Below are some general rules about segmentation.

1. Each utterance should be padded by a nominal 0.5 second buffer of silence on both sides. In general, these silence buffers can range from 0.35 to 0.75 seconds.
2. The boundary can **only** be placed in a “silence” consisting solely of channel noise and background noise. Whenever possible place the boundary in a section with very low energy (visually speaking, this is a flat part of the signal in the segmentation tool)
3. The 0.5 second buffers can contain breath noises, lip smacks, channel pops, and any other non-speech phenomena. However the boundary **can not** be placed in a noise of this sort.
4. No utterance can be longer than 15 seconds. As an utterance approaches 15 seconds in length, the validator is allowed to find a point of segmentation that will generate silence buffers less than 0.5 seconds but not less than 0.1 seconds. If this segmentation can not be found then that utterance should be marked as “NEEDS_REVIEW” in the log file and the validator should send an e-mail to the adjudication team explaining the problem.
5. Every utterance containing only silence must be greater than 1.0 seconds in duration.

6. Whenever possible choose a segmentation that maintains the phrase structure of the conversation. This means that, ideally, we would like every utterance to contain a single phrase. However, due to the nature of the SWB data, we realize that this is not always possible. **Note**: The previous instructions take precedence over this one.
7. The end of the preceding utterance coincides with the start of the next utterance. Hence all data is accounted for. Segmentation essentially involves placing a boundary between two utterances.
8. Consider a stretch of silence which has small amplitude noises embedded in it as a silence only utterance - do not mark the noise and do not segment the noises into separate utterances. However, if a noise has a particularly high amplitude, then segment it into its own utterance.

## 1.2. Transcription

1. Transcribe “verbatim,” without correcting grammatical errors: “i seen him,” “me and him gone to the movies,” etc.
2. Standard reductions and alternate pronunciations: Unless otherwise noted below, if “no” is meant but said as “naw” or “nah”, transcribe it how it is spoken. e.g. “y’all” instead of “you all”; “gonna” instead of “going to”; “wanna” instead of “want to”. However, in cases where there is severe reduction of a preposition such as in “kinda”, “sorta”, “gotta”, etc., transcribe the phrase as it was intended to be spoken. e.g. “kind of”, “sort of”, “got to”.
3. Follow the dictionary on hyphenating compounds in clear-cut cases. But “when in doubt, leave them out.”
4. Compound words: All compound words should be transcribed as one word when such a word exists in the dictionary unless there is an acoustical pause between the two words. e.g. “someone”, “everyday”, “cannot”, etc.
5. Try to avoid word abbreviations: Fort Worth, not Ft. Worth; percent, not %; dollars, cents, and so forth.
6. Contractions are allowed. e.g. “there’ll”, “it’s”, “can’t”, etc.
7. Capitalization: Use normal capitalization on proper nouns. Do not capitalize the beginning of the sentence. Titles should be capitalized using the standard grammar rule: the first word of a title is always capitalized, prepositions within a title that are under five letters are always lowercase, and the last word of a title is always capitalized.

   Example: “Dances with Wolves”, “Gone with the Wind”
8. The pronoun “I” should not be capitalized, instead it should be typed as “i”. Titles containing the word “I” are exceptions to this rule.

   Examples: i am tired of talking to you

   are you as tired as i am of listening to this
9. No punctuation should be used in the transcriptions.

10. Remember to watch for common spelling confusions like: its and it’s, they’re, there and their, by and bye, to and too, etc.

11. Numbers: Spell out all number sequences except in cases such as “123” or “101” where the numbers have a specific meaning. Transcribe years like 1983 as spoken — “nineteen eighty three.” Do not use hyphens (“twenty eight”, not “twenty-eight”).

12. Letter sequences: Spell out letter sequences: DFW, USA, FBI, NASA, ROM. When a letter sequence is used as part of an inflected word, add the inflection to the end of the letter sequence: e.g. TIer, BSing, the Oakland As, a witness IDed him. Transcribe a spoken spelling in all capital letters, each separated by a space: e.g. “dog is spelled D O G”; “my name is Tirelly, that’s T I R E L L Y”. If letter sequences contain special symbols then transcribe them as they would be written not as they are spoken: e.g. “AT&T” not “AT and T”; “Texas A&M” not “Texas A and M”.

13. Classifications of music are not titles, should not be transcribed in uppercase: e.g. “country western”, not “Country Western”; “rock ‘n’ roll”, not “Rock ‘n’ Roll”.

14. Possessives: Use standard grammar rules to denote possession: the US’s policy, Sally’s book, the drivers’ cars, the CEO’s decision, the dancers’ shoes.

15. Partial words: If a speaker does not completely pronounce a word and the word is not a standard reduction then spell out as much of the word as is pronounced, and inside brackets spell out the part of the word that was not pronounced. Use a single dash after the brackets if the last part of the word was not pronounced and a single dash before the brackets if the first part of the word was not pronounced to flag that a partial word was spoken. Context should be used to determine what word was intended to be spoken. If, from context, a reasonable intended word can not be determined, mark it as [vocalized-noise]

    Example: If a person begins to say the word “went” but only pronounces the “w”, transcribe it as “w[ent]-”.

    If a person says only the “at” portion of “that”, transcribe it as “-[th]at”.

16. Restarts of “i”: If a speaker restarts when saying the word “i”, it should be transcribed as “i-”. This should only be used when the first “i”s are not completely pronounced.

    Example: i- i really felt like i’ve been working now for about four years

17. Mispronunciations: If a speaker mispronounces a word and the mispronunciation is not an actual word, transcribe the word as it is spoken followed by the word that was intended. Divide these two words by a forward slash and enclose both words in brackets.

    Example: i wasn’t sure that they were blaming that [splace/space] space disaster on one company

18. Coinages: If a speaker uses and gives meaning to a word that is not an actual word, spell the word out as it sounds and enclose it in braces.

    Example: How are things for you {weatherwise}

19. Asides: If one of the speakers involved in the conversation talks to someone in the background and the words can be understood, then transcribe it as an aside enclosed in the markers, <b_aside> and <e_aside>. This only applies if one of the conversation speakers is involved in the background conversation. If just background speakers can be heard then this can be thought of either as noise or background noise depending energy level of the background speakers. compared to the foreground speakers.

    Example: “yeah i know what you <b_aside> honey i can’t play with you right now i’m on the phone <e_aside> sorry you know kids always want mommy all to themselves”

20. Hesitation sounds: Use “uh” or “ah” for hesitations consisting of a vowel sound, and “um” or “hm” for hesitations with a nasal sound, depending upon which transcription the actual sound is closest to. Use “huh” for the aspirated version of the hesitation as in: "huh? <other speaker responds> um ok, I see your point."

21. Yes/no sounds: Use “uh-huh” or “um-hum” (yes) and “huh-uh” or “hum-um” (no) for anything remotely resembling these sounds of assent or denial; you may use “yeah,” “yep,” and “nope” if that is what the words sound like.

22. Non-speech sounds during conversations: transcribe these using only the following list of expressions in brackets:

    [laughter] [noise] [vocalized-noise]

    Pick the closest description ([noise] will be adequate in most cases).

23. Laughter during speech: If laughter occurs directly before a word, place the [laughter] tag before the spoken word. If laughter occurs after a spoken word, place the [laughter] tag after the word. If the speaker laughs while saying the word, but the word is still understood, transcribe this as [laughter-word], where "word" is the word spoken during the laughter. If the speech is obliterated by the laughter, transcribe it strictly as [laughter]. If a speaker laughs while saying several words and the words are understood, transcribe each word in the phrase as [laughter-word]. Laughter throughout the phrase, “you don’t say,” would be transcribed as: [laughter-you] [laughter-don’t] [laughter-say].

24. Pronunciation variants: The convention of "word_1" is used to denote a common variation in the pronunciation of a word. A list of these words will be kept in the transcription conventions documentation. Examples of pronunciation variants currently in use are:

    | | | | |
    |---|---|---|---|
    | about_1 | b aw t | because_1 | k ah z |
    | depends_1 | p eh n d z | them_1 | eh m |
    | okay_1 | m k ey | especially_1 | s p eh sh ax l iy |

    These are to be used judiciously, and only to capture frequently occuring reductions which are easy to distinguish from the baseform.<

25. Continuous background noise: Consider it as part of channel. For example, if a baby cries at a consistent energy level throughout the conversation then treat it as background

noise. Only consider it as noise if the noise grows much louder than the normal level — in our example above the baby screaming would warrant considering it as noise. In this case mark it as [noise].

26. Special lexicon issues:

- Use "all right" instead of “alright” in all cases.

- Use "Walkman" when the speaker is referring specifically to the Sony Walkman, and use “walkman” when there is no reference to Sony.

  Example: i like to listen to my walkman when exercising
  i wonder how many transistors a Sony Walkman has?

- Use “doggy” instead of “doggie” in all cases.

- Use "God" instead of "god" in all cases.

  Example: it’s like you know God what are they doing

## 2. Original Instructions

Following is the original set of guidelines and instructions for transcription of SWITCHBOARD. We propose to deviate from these in a manner explained previously in Section 1.

### 2.1. General Instructions

1. Transcribe “verbatim”, without correcting grammatical errors: "i seen him,” “me and him gone to the movies,” etc.
2. Do not try to imitate pronunciation; use a dictionary form: “no” will do for “naw,” “nah,” etc., “oh” for “aw,”; “going to” (not gonna or goin to); “you all” rather than “y’all”; “kind of” instead of “kinda”; etc. Nonstandard words which are not in the dictionary (e.g., kiddo) should be typed normally, i.e. without quotes or other special notation.
3. Follow the dictionary on hyphenating compounds in clear-cut cases. But “when in doubt, leave them out.”
4. Try to avoid word abbreviations: Fort Worth, not Ft. Worth; percent, not %; dollars, cents, and so forth.
5. Contractions are allowed, but be conservative. For example, contraction of “is” (it’s a boy, running’s fun) is common and standard, but there’ll (there will) be forms that’re (that are) better left uncontracted. It is always permitted to spell out forms in full, even if the pronunciation suggests the contracted form. Thus it is O K to type he is and they are and we would even if it’s he’s and they’re and we’d you heard.
6. Use normal capitalization on proper names of persons, streets, restaurants, cities, states, etc., but put titles (of books, journals, movies, songs, plays, TV shows, etc.--what would properly be in italics.) in ALL CAPS, i.e., uppercase letters.
7. If it is necessary to use accent marks, insert the number 3 before the letter which would receive the accent, e.g., fianc3e.
8. Punctuation: although normal punctuation rules apply, spontaneous conversational speech is full of difficult situations. Strive for simplicity and consistency, with the following specific guidelines:
    - terminate each sentence with a period unless a question mark or exclamation point is clearly justified;
    - use a comma instead of ... or -- or fancier punctuation when speakers change thoughts or grammatical structures in the middle of a sentence;
    - for more detail, and for special rules involving interruptions, etc., see below under SPECIAL CONVENTIONS.
9. Be sure to run a spell check upon completion of the transcript. Remember to watch for common spelling confusions like: its and it’s, they’re and there and their, by and bye, etc.

[TV]
[baby]
[baby_crying]
[baby_talking]
[barking]
[beep]
[bell]
[bird_squawk]
[breathing]
[buzz]
[buzzer]
[child]
[child_crying]
[child_laughing]
[child_talking]
[child_whining]
[child_yelling]
[children]
[children_talking]
[children_yelling]
[chiming]
[clanging]
[clanking]
[click]
[clicking]
[clink]
[clinking]
[cough]
[dishes]
[door]
[footsteps]
[gasp]
[groan]
[hiss]
[horn]
[hum]
[inhaling]
[laughter]
[meow]
[motorcycle]
[music]
[noise]
[nose_blowing]
[phone_ringing]
[popping]
[pounding]
[printer]
[rattling]
[ringing]
[rustling]
[scratching]
[screeching]
[sigh]
[singing]
[siren]
[smack]
[sneezing]
[sniffing]
[snorting]
[squawking]
[squeak]
[static]
[swallowing]
[talking]
[tapping]
[throat_clearing]
[thumping]
[tone]
[tones]
[trill]
[tsk]
[typewriter]
[ugh]
[wheezing]
[whispering]
[whistling]
[yawning]
[yelling]

Table 1. A list of typical non-speech sounds that are transcribed as “[noise]”. Effort expended on extremely detailed marking of noise has not proven productive to date.

## 2.2. Special Conventions for SWITCHBOARD Conversations

1. Speakers should be indicated by “A:  ” and “B:  ” at the left margin, with two spaces after the colon, and with a blank line between speakers (i.e., an extra carriage return before each A: or B:). On the audio tape, A will be THE SPEAKER ON THE FIRST OF THE TWO SEPARATELY RECORDED SIDES. IT IS IMPERATIVE TO KEEP THIS DESIGNATION CORRECT AND CONSISTENT, even when the crosstalk or echo is so strong that both speakers are equally loud. The log sheet for each conversation will show the first few words by each speaker, to help you confirm the assignment.

   EXAMPLE:
   A:  Blah blah blah blah.
   B:  Blah blah blah.
   A:  Etcetera.

2. Spell out letter and number sequences: D F W, seven forty-seven, US A, one oh one, F B I, etc., unless the letter sequence is pronounced as a word, as in NASA, ROM, DOS.

3. Transcribe years like 1983 as “nineteen eighty-three,” with hyphens only between the tens and ones digits.

4. When a letter sequence is used as part of an inflected word, add the inflection with a dash: T I -er, B S -ing, the Oakland A -s, a witness I D -ed him. This leads to

clumsy-looking possessive forms, as in: the U S -'s policy, the T I -er's last name, all the C E O -s' votes, but it saves lots of time later on.

5. Partial words: if a speaker does not finish a word, and you think you know what the word was, you may spell out as much of the word as is pronounced, and then use a single dash followed by a comma, -,. If you cannot tell what word the speaker is trying to say, leave it out.

   EXAMPLE:
   A: Well, th-, that's what they kept tell-, wanted me to believe.
   B: I, I, I just am not to-, totally sure, uh, about that.

6. Hesitation sounds: use "uh" for all hesitations consisting of a vowel sound (rather than trying to distinguish uh, ah, er, etc.), and "um" for all hestitations with a nasal sound (rather than uhm, hm, mm, etc.)

7. Yes/no sounds: use "uh-huh" (yes) and "huh-uh" (no) for anything remotely resembling these sounds of assent or denial; you may use "yeah," "yep," and "nope" if that is what the words sound like.

8. Punctuation: use commas instead of ... or -- or other "fancy" punctuation when speakers change thoughts or grammatical structures in the middle of a "sentence." Terminate each sentence with a period unless a question mark or exclamation point is clearly justified. Only use suspension dots ... if a speaker leaves a sentence unfinished at the end of his/her turn, and a period cannot be used, or at the end of a conversation where the speaker's turn was cut off by the computer timing out:

   EXAMPLE:
   A: I was going to do that, but then I ...
   B: Right, me too.

9. Use a double dash if a speaker breaks a sentence off and picks it up at the beginning of the next turn, with another double dash where the pickup begins:

   EXAMPLE:
   A: I was going to do that, but then I --
   B: Right, me too.
   A: -- thought I better not after all.

10. Non-speech sounds during conversations: indicate these using only the following list of expressions in brackets. When making judgments, pick the closest description; [noise] will be adequate to describe most sounds that are not represented below in Table 1. Note underscores (not spaces or hyphens) to connect the double word descriptions.

11. If the event being described lasts longer than a few words, then indicate the beginning in brackets [ ], and the end in brackets with a "/", [/].

    EXAMPLE:

1. Separate multiple sounds by a space, each one in brackets:

A: Oh, that’s funny. [laughter] [cough] Excuse me, I have a cold.
B: That’s all right, [sneezing] so do I. [barking] [child_talking]

2. Use “/” to show end of a continuous sound:

A: Well, it all depends, uh, on, you know, [baby_crying] how the family reacts. I mean, it can be a positive or a negative thing, you know?
B: Yeah, well, I guess so. It just seems [/baby_crying] to me that it’s a very difficult, uh, difficult issue.

12. When a comment is needed to describe an event, put the comment in curly braces { }: {very faint}, {sounds like speaker is talking to someone else in the room}, {speaker imitates a woman’s voice here}.

EXAMPLE:

1. Curly braces to describe the speech:

B: Yeah, yeah, I agree {very faint} right.

2. Combine curly braces and brackets if more explanation is needed to describe the word in the brackets:

A: Did it sound like this? [clicking] {sounds made with mouth}
B: No, more like [clicking] {sounds like a pencil tapping on a table} this.

13. When a word or phrase is not clear, type DOUBLE PARENTHESES (( )) around what you think you hear. If there is no way to tell what the speaker said, leave 1 blank space between the double parentheses, indicating speech has been left out because it was unintelligible.

EXAMPLE:

A: So when I finally did ((take up)) the violin, progressed pretty quickly in the beginning.
B: Of course, that was in college which was a long time ago, so (( )) I remember.

14. Marking untopical speech for possible trimming: Use an “at sign”, @, and a double “at sign”, @@, to designate potential “trim points” at the beginning or end of conversations. These would exclude speech that either is not part of the conversation itself, or refers directly to the protocol. For example, it sometimes happens that callers accidentally press the touchtone button that begins recording, and are being recorded during the “warmup period” and don’t know it. All such speech should be marked for trimming. Other examples would be speech that:

a) explicitly refers to the SWITCHBOARD protocols;
b) refers to the process of making the call;
c) uses the TITLE of the prompt (e.g., “music”); or
d) repeats or paraphrases the PROMPT itself.

15. [The TITLE and the PROMPT for each topic will be found on your information sheet; they are keyed to the topic number, which is on the log sheet for each conversation.]

16. Marking these trim points means that EVERYTHING BEFORE '@' AND/OREVERYTHING AFTER '@@' may be discarded without losing the main body of the conversation on the topic. These symbols may therefore only be used ONCE AT THE BEGINNING (@) AND/OR ONCE AT THE END (@@) of the conversation. They must also be used ONLY AT TURN-TAKING POINTS, i.e., at the left hand margin, before an "A:" or "B:", NOT part of the way through someone's turn. One or both may be used in a single conversation, i.e., trimming of material at the beginning is independent of trimming at the end.

17. Social niceties and transitional talk are neutral. That is, they may be left alone, but should be trimmed if they occur next to material that definitely deserves trimming.

EXAMPLE:

A: Okay, so what am I supposed to do now? Wait, let me read,
B: I think you're supposed to push one.
A: let's see, it says here to push, okay, but I think I already,
okay are you ready?
B: Yep. [Talking about protocol up to here.]
A: Here we go. Alright, now, tell me, what is your favorite kind
of music? [Using topic TITLE explicitly.]
@B: I enjoy Mozart and reggae, but I really love rap. [OK]
.
. <body of conversation is here>
.
A: I've certainly enjoyed hearing what you have to say. [Trim optional here.]
@@B: Well, if we've talked enough, do I need to push a button or anything? I guess not, we can just hang up. So long. [Talk of protocol should be trimmed.]
A: Bye. Nice talking to you.

ANOTHER EXAMPLE:

A: Hi, there, how are you doing?
B: Fine, how about you?
A: Just great, except for all this heat. [Chitchat up to here could be left alone if no reason to trim occurred.]
B: Well. Care of the elderly, huh? That's our topic? [Need to trim because it mentions the topic TITLE.]
@A: Yes. Do you have any relatives that need special care? [This is OK as part of the conversation, since only the word "care" is repeated from the prompt. It is not trimmed--initial trimming ends with the '@'.]
.
.
.
@@B: Well, I guess we have solved the problem of care of the elderly, and

how to choose nursing homes, haven't we? [Trimmed because it contains both TITLE and a paraphrase of prompt.]

A: Sure did. I hope your grandmother gets better. So long now, it's been fun talking to you. [Social pleasantries would not be trimmed themselves, but no harm in trimming them in order to get rid of the previous turn.]

18. Simultaneous talking: Wherever possible, mark where both speakers talked simultaneously with TWO PAIRS of pound signs (#), ONE BEFORE AND ONE AFTER each of the segments spoken at the same time. One of these segments MUST BEGIN A TURN; in other words, if one person is an "interruptor", his interruption starts a new turn. Remember, BOTH speakers' turns must contain TWO pound signs each.

A SIMPLE EXAMPLE:

A: Okay, well, I guess that's about it.
B: Yeah.
A: Nice talking to you.
B: # Right, bye. #
A: # Bye bye. #

ANOTHER EXAMPLE:

A: I never heard such nonsense, you know,
B: # Yeah, I know. # [B interrupts while A continues.]
A: # as I heard that # day when I blah blah blah. [A continues beyond the simultaneously spoken words.]

WHICH COULD ALSO BE WRITTEN:

A: I never heard such nonsense, you know, # as I heard that #
B: # Yeah, I know. #
A: day when I blah blah blah

ANOTHER EXAMPLE:

A: I never heard such nonsense, # you know, # [A starts.]
B: # Yeah, # [B starts to step on A.]
A: as I heard that day when # I was at that meeting. # [A continues without stopping.]
B: # I agree with you all the way # [B comes in over A again.]