The 1997 Hub-5NE Evaluation Plan for Recognition of Conversational Speech over the Telephone, in Non-English Languages

Last Modification: August 19th, 1997
Version: 1.0

Introduction

The 1997 Hub-5NE evaluation is part of an ongoing series of periodic 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 conversational speech recognition. To this end the evaluation was designed to be simple, to focus on core speech technology issues, to be fully supported, and to be accessible.

The Hub-5NE evaluation, conducted in the fall, complements another related evaluation which is conducted in the spring. The spring evaluation focuses on the recognition of conversational speech in English. This evaluation is dedicated to the advancement of speech recognition technology for languages other than English, and specifically this year for Arabic, German, Mandarin, and Spanish. It focuses also on issues related to porting recognition technology to new languages, to system generality, and to language commonalties and universals.

The 1997 Hub-5NE evaluation will be conducted in September. (Data will go out on August 29, and results are due back by September 19.) A follow-up workshop for evaluation participants will be held November 4-6 to discuss research findings. Participation in the evaluation is solicited for all sites that find the task worthy and the evaluation of interest. For more information, and to register a desire to participate in the evaluation, please contact Dr. Alvin Martin at NIST.[1]

Technical Objective

The Hub-5E evaluation focuses on the task of transcribing conversational speech into text. This task is posed in the context of conversational telephone speech in Arabic, German, Mandarin, and Spanish. The evaluation is designed to foster research progress, with the goals of:

1. exploring promising new ideas in the recognition of conversational speech,

2. developing advanced technology incorporating these ideas, and

3. measuring the performance of this technology

The Task

The task is to transcribe conversational speech. The speech to be transcribed is presented as a set of conversations collected over the telephone. Each conversation is represented as a "4-wire" recording, that is with two distinct sides, one from each end of the telephone circuit. Each side is recorded and stored as a standard telephone codec signal (8 kHz sampling, 8-bit mu-law encoding).

Each conversation is represented as a sequence of "turns", where each turn is the period of time when one speaker is speaking. Each successive turn results from a reversal of speaking and listening roles for the conversation participants. The transcription task is to produce the correct transcription for each of the specified turns. The beginning and ending times of each of these turns will be supplied as side information to the system under test. This turn information will be supplied in

PEM[2] format, with one PEM file for all conversations to be transcribed. (Note that the turns are not necessarily a simple sequence of non-overlapping time intervals. They may be overlapping or non-alternating from time to time, because there is no sequencing constraint on conversational interaction.)

Speech Data

Training Data

All of the Call_Home data in each language that is designated as training data may be used. In addition, in Arabic, Mandarin and Spanish the 20 conversations originally designated as development data may be used for training. In Mandarin and Spanish, the 1995 evaluation data may also be used for this purpose. All of this data is available from the LDC. Additional data from other corpora may also be used for training, provided that the data is made publicly available at the time of reporting results.

Development Data (the DevSet)

In Arabic, Mandarin, and Spanish the 1996 evaluation set of 20 conversations will serve as the development data for the present evaluation. In German the 20 conversations designated as development data will be used.

Evaluation Data (the EvalSet)

The EvalSet will comprise 20 Call_Home conversations for each language. Whole conversations will be supplied, but recognition will be scored only for a 5 minute excerpt chosen from each conversation. Speaker turn segmentation information for these 5 minute excerpts will be supplied to guide the recognition system. This segmentation information will be supplied in NIST's PEM file format.

The Evaluation

Each system will be evaluated by measuring that system's word error rate (WER), except in Mandarin, where character error rate (CER) will be the primary error measure. Each system will also be evaluated in terms of its ability to predict recognition errors. System performance will be evaluated over an ensemble of conversations. These conversations will be chosen to represent a balanced sampling from within the available pool of conversations of conditions of evaluation interest that have indicated by the transcribers. These will include sex, age, rate of speech, background and channel noise, and difficulty of transcription

The Reference Transcription

The reference transcriptions are intended to be as accurate as possible, but there will necessarily be some ambiguous cases and outright errors. In view of the existing high error rates of automatic recognizers on this type of data, it is not considered cost effective to generate multiple independent human transcriptions of the data or to have a formal adjudication procedure following the evaluation submissions.

The reference transcription for each turn will be limited to a single sequence of words. This word sequence will represent the transcriber's best judgment of what the speaker said.

Further details are explained in the "Scoring Issues" section below.

The WER (CER) Metric

Word error rate is defined as the sum of the number of words in error divided by the number of words in the reference transcription. The words in error are of three types, namely ***substitution*** errors, ***deletion*** errors, and ***insertion*** errors. Identification of these errors results from the process of mapping the words in the reference transcription onto the words in the system output transcription. This mapping is performed using NISTËs SCLITE software package[3].

- A substitution error results when the spellings of the reference word and the corresponding system output word differ.
- A deletion error results when the reference word has no corresponding system output word.
- An insertion error results when a system output word has no corresponding reference word.

Scoring will be performed by aligning the system output transcription with the reference transcription and then computing the word error rate. Alignment will be performed independently for each turn, using NIST’ s SCLITE scoring software. The system output transcription will be processed to match the form of the reference transcription.

For Mandarin, character error rate alignment and scoring will be performed similarly, but at the character level.

Scoring Issues

The reference transcription will be transformed prior to comparing it with the output from a recognizer. It is important that these transformations are properly comprehended in the design of a recognition system, so that the system will perform well according to the scoring measure. Here are the transformations that will be applied to the reference:

Word fragments

Word fragments are represented in the transcription by appending a "-" to the (partial) spelling of the fragmented word. Fragments are included in the total word count and scored as follows:

1) If the fragment is deleted in the time alignment process, no error is counted

2) If the fragment matches the recognizer output up to the "-", no error is counted

3) Otherwise, there is a substitution error

Unintelligible and Doubtful Words

The reference transcripts may describe some speech as unintelligible (indicated by "((   ))"), and then may or may not also provide a "best guess" as to what words it consists of. Such "best guess" doubtful words will be included in the total word count, with scoring as follows:

1) If the alignment produces a deletion, no error is counted

2) If the alignment produces a matching word, no error is counted

3) Otherwise, there is a substitution error

Foreign Words

The reference transcripts may describe words as foreign, as words not in the language under test. This description will not be applied to words of foreign origin that have been widely incorporated into speech of the given language. Such foreign words will be included in the total word count, with scoring as follows:

1) If the alignment produces a deletion, no error is counted

2) If the alignment produces a matching word, no error is counted

3) Otherwise there is a substitution error

Pause fillers

For scoring purposes, all hesitation sounds, referred to as "non-lexemes", will be considered to be equivalent, and will be scored the same way as fragments, doubtful, and foreign words. Although these sounds are transcribed in a variety of ways due to highly variable phonetic quality, they are all considered to be functionally equivalent from a linguistic perspective. Thus, all reference transcription words beginning with "%", the hesitation sound flag, along with the conventional set of hesitation sounds, will be mapped to "%hesitation". The system output transcriptions should use any of the hesitation sounds (without "%") when a hesitation is hypothesized or omit it altogether. Again:

1) If the alignment produces a deletion, no error is counted

2) If the alignment produces a matching word, no error is counted

3) Otherwise there is a substitution error

The evaluation data distributed for each language will contain the list of recognized hesitation sounds for the language.

Multiple spellings

Some words appear in the training corpus with multiple spellings, including misspellings. For scoring, however, a single standardized spelling will generally be required, and the recognizer must output this standard spelling in order to be scored as correct. The evaluation data distributed for each language will list the allowed alternate word spellings, if any, for the language.

Homophones

Homophones will not be treated as equivalent. Homophones must be correctly spelling in order to be counted as correct.

Overlapping speech

Periods of overlapping speech will not be scored. Any words hypothesized by the recognizer during these periods will not be counted as errors.

Compound Words

Compound words will be treated as separate multiple words if they commonly appear in that form (in the training data or official lexica). If a compound word exists only in compound form, only then will it be treated as a single word.

Contractions

Contractions will be expanded to their underlying forms in the reference transcriptions. Manual auditing will be used to ensure correct expansion. Contractions in the recognizer output will be expanded based on default expansions for standard contractions in the language. Thus the recognizer need not expand contractions, but it may be preferable for it to do so.

Language Specific Issues

Arabic: The definite article will be detached from the following word before scoring.

German: Compound words are frequent and often include multiple words. The compound word rules will apply. Thus scoring will not be affected by whether or not compounds are broken into separate words. Furthermore, the scoring software will identify and divide compound words even when spelling changes occur as a result of compounding.

Mandarin: Character error rate, rather than word error rate, will be the performance measure, as noted above. Furthermore, confidence measures, as discussed below, will be applied at the character level. If the system output gives confidences only at the word level, the word level values will be automatically imputed to characters making up the word.

Spanish: In past evaluations in Spanish, a lexeme error rate was used in addition to word error rate. This will *not* be used in this evaluation.

The Confidence Measure

Along with each word output by a system, a confidence measure is also required. This confidence measure is the system's estimate of the probability that the word is correct. (In Mandarin, the confidence may be specified for character. If confidences are specified at the word level, the word values will be imputed to each character of a word.) While this might be merely a constant probability, independent of the input, certain applications and operating conditions may derive significant benefit from a more informative estimate that is sensitive to the input signal. This benefit will be evaluated by computing the mutual information (cross entropy) between the correctness of the system's output word and the confidence measure output for it, normalized by maximum cross entropy:

$$\text{Confidence\_Score} = \left\{ H_{max} + \sum_{correct\,w} \log_2\left(\hat{p}(w)\right) + \sum_{incorrect\,w} \log_2\left(1 - \hat{p}(w)\right) \right\} / H_{max},$$

$$\text{where } H_{max} = -n \log_2\left(p_c\right) - \left(N - n\right)\log_2\left(1 - p_c\right),$$

n = the number of correct HYP words,

N = the total number of HYP words,

p_c = the average probability that an output word is correct $= n / N$, and

$\hat{p}(w)$ = the confidence measure output, as a function of output word.

Submission of Results

Results must be submitted to NIST by September 19, 1997 at 1:00 pm. EST using the following steps:

1. system output file creation,
2. directory structure creation,
3. system documentation, *including execution times,* and system output inclusion
4. transmission protocol to NIST.

Step 1: System output File creation

The time-marked hypothesis words for each (single language) test will be placed in a single file, called "<TEST_SET>.ctm". The CTM (Conversation Time-Mark) file format, is a concatenation of time marks for each word in each side of a conversation. Each word token must have a conversation id, channel identifier [A | B], start time, duration, case-insensitive word text, and a confidence score. The start time must be in seconds and relative to the beginning of the waveform file. The conversation id's for this evaluation will be of the form:

> CONV_ID::= ll_DDDD (where ll is a two letter lower-case language code and where DDDD is a four digit conversation code)

For the Mandarin evaluation, sites that choose to supply confidence scores at the character level must create a separate CTM record for each character. Otherwise, confidence scores for multi-character words will be imputed to all characters.

The file must be sorted by the first three columns: the first and the second in ASCII order, and the third by a numeric order. The UNIX sort command: "sort +0 -1 +1 -2 +2nb -3" will sort the words into appropriate order.

Lines beginning with ';;' are considered comments and are ignored. Blank lines are also ignored.

Included below is an English example:

```
;;
;; Comments follow ';;'
;;
;; The Blank lines are ignored

;;
en_7654 A 11.34 0.2 YES -6.763
en_7654 A 12.00 0.34 YOU -12.384530
en_7654 A 13.30 0.5 CAN 2.806418
en_7654 A 17.50 0.2 AS 0.537922
:
en_7654 B 1.34 0.2 I -6.763
en_7654 B 2.00 0.34 CAN -12.384530
en_7654 B 3.40 0.5 ADD 2.806418
en_7654 B 7.00 0.2 AS 0.537922
:
```

Step 2: Directory Structure Creation

Create a directory identifying your site ('SITE') from the following list which will serve as the root directory for all your submissions:

- bbn
- dragon
- ibm
- cmu
- .
- .
- .

You should place all of your recognition test results in this directory. When scored results are sent back to you and subsequently published, this directory name will be used to identify your organization.

For each test system, create a sub-directory under your 'SITE' directory identifying the system's name or key attribute. The sub-directory name is to consist of a free-form system identification string 'SYSID' chosen by you. Place all files pertaining to the tests run using a particular system in the same SYSID directory.

The following is the BNF directory structure format for Hub-5NE hypothesis recognition results:

<SITE>/<SYSID>/<FILES>

where,

SITE ::= bbn | dragon | ibm | sri | . . .
SYSID ::= (short system description ID, preferably <= 8 characters)
FILES ::=

sys-desc.txt

(system description, described below, including reference to paper if applicable)

<TEST_SET>.ctm

(file containing time-marked hypothesis word strings created in Step 1) where, TEST_SET ::= arabic|german|mandarin|spanish

Step 3: System Documentation, *including execution times,* and System Output Inclusion

For each test you run, a brief description of the system (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. In this case, however, the submitting site must identify one system as the "primary" system prior to performing the evaluation.)

The format for the system description is as follows:

SITE/SYSTEM NAME
TEST DESIGNATION

1. Primary Test System Description:
2. Acoustic Training:
3. Grammar Training:
4. Recognition Lexicon Description:
5. Differences for each Contrastive Test: *(if any contrastive test were run.)*
6. New Conditions for This Evaluation:
7. Execution Time:

   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.

8. References:

Your system description file should be placed in the 'SYSID' sub-directory which it pertains to and must be called, "sys-desc.txt".

Likewise, the time-marked hypothesis file, created in step 1, should be placed in the 'SYSID' sub-directory which it pertains to, and must be called, "<TEST_SET>.ctm". For this evaluation, the value for <TEST_SET> will be the language name, i.e. "arabic" or "german" or "mandarin" or "spanish".

Step 4: Test Results Submission Protocol

Once you have structured all of your recognition results according to the above format, you can then submit them to NIST. Due to international e-mail file size restrictions, test sites are permitted to submit results to NIST using either email or anonymous ftp. Continental US sites may use either method, but international sites must use the 'ftp' method. The following instructions assume that you are using the UNIX operating system. If you do not have access to UNIX utilities or ftp, please contact NIST to make alternate arrangements.

E-mail method:

First change directory to the directory immediately above the <SITE> directory. Next, type the following:

```
tar -cvf - ./<SITE> | compress | uuencode <SITE>-<SUBM_ID>.tar.Z | \
mail -s "September 97 Hub-5NE test results <SITE>-<SUBM_ID>" \
alvin.martin@nist.gov
```

where,

<SITE>

is the name of the directory created in Step 2 to identify your site.

<SUBM_ID>

The submission number (e.g. your first submission would be numbered '1', your second, '2', etc.)

Ftp method:

First change directory to the directory immediately above the <SITE> directory. Next, type the following command.

```
tar -cvf - ./<SITE> | compress > <SITE>-<SUBM_ID>.tar.Z
```

where,

<SITE> is the name of the directory created in Step 2 to identify your site. <SUBM_ID> The submission number (e.g. your first submission would be numbered '1', your second, '2', etc.)

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

- ftp> cd /pub/benchmark/sep97_hub-5ne
- ftp> binary
- ftp> put <SITE>-<SUBM_ID>.tar.Z
- ftp> quit

You've now submitted your recognition results to NIST. The last thing you need to do is send an e-mail message to Alvin Martin at 'alvin.martin@nist.gov' notifying NIST of your submission. Please include the name of your submission file in the message.

Note:

If you choose to submit your results in multiple shipments, please submit ONLY one set of results for a given test system/condition unless you've made other arrangements with NIST. Otherwise, NIST will programmatically ignore duplicate files.

Schedule

Commitment Deadline	August 22, 1997
EvalSet Release	August 29, 1997
Results Deadline	September 19, 1997 at 1:00 pm. EST
Workshop	November 4-6, 1997 Maritime Institute of Technology and Graduate Studies Linthicum, Maryland

Footnotes

1. 1. To contact Dr. Martin, you may send him email at Alvin.Martin@nist.gov, or you may call him at 301/975-3169.

2. The PEM ("partitioned evaluation map") file format is given in the SCLITE documentation available through NIST's web page (http://www.nist.gov/speech/software.htm). Each record contains five fields: <filename>, <channel ("A" or "B")>, <speaker ("unknown">, <begin time> and <end time>.
3. SCLITE software is available via NIST's web page (http://www.nist.gov/speech/software.htm).