INITIAL DRAFT GUIDELINES FOR THE DEVELOPMENT OF THE NEXT-GENERATION SPOKEN LANGUAGE SYSTEMS SPEECH RESEARCH DATABASE

George R. Doddington,
TI Senior Fellow and Chief Speech Scientist
Computer Science Center
Texas Instruments

OBJECTIVE

To best serve the strategic needs of the DARPA SLS research program by creating the next-generation speech database(s).

- To promote progress on the important SLS research problems:
    - phonetic modeling (acoustic-phonetic decoding)
    - higher level modeling of the speech mechanism
    - language modeling (above the level of speech)
- To be adequate, practical, and timely
    - comprehensive enough to support true learning
    - limited enough to be accomplished
    - soon enough to be valuable

BACKGROUND

The DARPA speech research program has made significant progress in the development of speech recognition technology. This progress has led to an increase in the overall scope of the program, from "speech recognition" to "spoken language recognition". This change in the nature of the research effort requires that a new speech database be developed to better support the new research objectives.

The value of establishing a few well-designed databases, shared by all of the research contributors, was demonstrated in the first phase of the DARPA speech research program. Such databases integrate the various research efforts by providing a common research medium which affords a shared understanding of the problem and by self-direction of technical efforts through unequivocal demonstrations of the strengths/weaknesses of competing approaches.

DATABASE REQUIREMENTS

- Primary requirements:
    - "natural" (task-oriented) speech, unconstrained by fixed vocabulary or fixed syntax
    - support for all speech recognition problem areas
    - emphasis on speaker independent recognition

- Implied requirements:
  - -- comprehensive
  - -- large vocabulary
  - -- unconstrained -- large vocabulary, connected speech, no formal language specification
  - -- support for performance measurement
  - -- relevant to DoD needs

- Efficiency:
  - -- low cost per second of speech
  - -- high value per second of speech -- emphasis on problem areas
  - -- rapid development cycle

- Amount of speech:
  - -- 100 hours

- Number of speakers:
  - -- 1000 speakers

- Data documentation:
  - -- orthographic transcription
  - -- speaker description

The desiderata for the next-generation spoken language system speech database include the following:

- First, the speech must be "natural". This implies that the speech be spontaneous as well as unconstrained by vocabulary or syntax.

- Second, the speech should involve or simulate interactive man/machine problem solving. This implies that the speaker's focus of attention is not on the speech act itself but rather on the task which the speech serves.

- Third, the database should be sufficiently representative and general so that it can support SLS technology development that is useful in general, beyond the specific task domain of the database.

- Fourth, the database development effort must be doable, with the data available to the researchers soon enough to drive the research effort.

- Fifth, the database should be relevant to DARPA needs and should demonstrate a high intrinsic value of spoken language systems for DoD applications.

There is one other important consideration in the creation of an SLS database. The database should not be so difficult that it discourages the enthusiasm so essential for strong effort and steady progress in SLS technology research and development. This will be difficult for a spontaneous SLS database. Perhaps the database might be graded into categories of difficulty, thus allowing a gradual progress toward solutions for the most difficult problems.

ALTERNATIVE SOURCES OF DATA

- "Read" speech
  - -- Naval Battle Management
  - -- Acoustic-phonetic database (TIMIT)

- "Performance task" simulation, task-oriented naturally elicited (TONE)
  - -- NOSC naval battle management
  - -- Personnel database
  - -- Spread sheet
  - -- Other less well-defined task domains
    - -- Pilot's associate
    - -- Search & Rescue
    - -- Navigation Aid by Phone

- "Natural task" data collection
  - -- ATC operations
  - -- Travel planning and reservations
  - -- Theater recordings
  - -- Dictation
    - -- office correspondence
    - -- medical transcriptions

From among the three general categories of speech data, namely "read" speech, "performance task" speech, and "natural task" speech, probably the most satisfactory candidate is "natural task" speech. A discussion of the problems with the different categories of speech databases follows:

1) "Read" speech database problems: The principal objection to "read" speech data is that it does not support the development of technology to recognize natural spoken language. The argument that transcripts of spoken language could be read is weak. It is not even clear that a reasonable representation of natural speech could be presented to the subjects to be read. Most agree that "read" speech is not suitable, and so I will not argue further against this category.

2) "Performance task" simulation problems: Several factors contribute to make "performance task" simulation impractical. First, the effort to create a meaningful simulation would be difficult at best. An affordable effort could not support the creation of a target application, and would necessarily be limited to creating a simulation of a spoken language interface. But working within existing target applications will limit the effective language domain, because few if any existing applications have the large domain and vocabulary that the SLS effort proposes to study. How to support a large meaningful language with this simulation is a critical challenge. Second, there would be a large effort required to create meaningful task scenarios for the subjects, to find (or train) a sufficient number of knowledgeable subjects, and to orient and interest the subjects in the specifics of the assigned task. Third, the need to use a trained and skilled human translator (to listen to the subjects and input appropriate commands to the system simulation), will limit the amount of data that can be collected to a level that will not supply the needed amount of data for productive research.

3) "Natural task" speech database problems: There are several negative aspects to a "natural task" speech database, mostly related to the difficulty of the recognition task: Because of the in situ nature of the speech, there will be significant additional dimensions of variation in the speech signal. These will include such things as greater acoustic variation, and less or no conscious control of a subject on his speech. Further, the usage of vocabulary and syntax will be highly skewed, with rare forms occurring just often enough to maximize the error rate through insufficient exposure to training data. (All spontaneous speech databases will be subject to such skewing.)

The benefits of a "natural task" database, however, may outweigh or overcome the shortcomings. Most important, a natural task database allows study of the underlying principles that govern natural spoken language without the risk of corruption or bias of the language which might be caused by artifacts induced by simulation or data collection constraints. Also, the "natural task" scenario will support a higher level language, in terms of intelligent command/control dialog, than could be expected from a database query system such as spread sheet or personnel database. Finally, a most important benefit of the "natural task" scenario is that it can much more efficiently provide a large speech database. This can help to overcome the skewed distribution of vocabulary and syntax.

Of the candidate tasks listed above, the first two, ATC operations and travel planning, probably fit the SLS database objectives best. They are both highly interactive and relatively well controlled. (ATC controllers use head-mounted microphones, and travel planning uses the telephone.) Also, they are both highly interactive and, at least in principle, support easy and efficient speech data collection. (ATC already record their data, although the bandwidth is typically limited to less than 4 kHz.) Dictation would be a good candidate, except that it is really not an interactive speech task, and therefore it does not support the SLS objective directly.

EVALUATION ISSUES

- Development/evaluation partition

- Performance definition
  % phones correctly recognized
  % words correctly recognized
  % sentences correctly recognized
  % sentences correctly understood
  % tasks correctly accomplished
  % increase in productivity

- Error classification --
  -- acoustic (simple decoding error)
  -- lexical (word not in lexicon)
  -- syntactic (ungrammatical or insufficient grammar)
  -- semantic (misunderstood)

In progressing from "speech recognition" to "spoken language systems", two new dimensions will be added to performance evaluation. First, "coverage" will be an important, perhaps dominant, factor in system performance. Although this has always been an important issue in the practical use of speech recognition, it has typically been ignored, and evaluation has considered only speech data that falls within the formal language model. Second, it will no longer be adequate to evaluate speech at the level of orthographic transcription. It will be necessary to "understand" the speech sufficiently to determine the appropriate response to it. This is the greatest strategic challenge facing the SLS research effort, and one for which there will be no general near-term solution. The hope is that technology may be developed to support specific task-domain applications. The immediate challenge is to define an evaluation methodology for spoken language understanding.

IMPORTANT PROBLEM DIMENSIONS IN THE DATABASE

- Acoustics
  -- microphone
  -- noise
  -- channel

- Speaker
  -- sex
  -- other physiological variables
  -- speech patterns

- Accent/dialect
- Domain
  task
  lexicon
  dialog

RELEVANCE OF CURRENT DATABASES (STRENGTHS/WEAKNESSES/LESSONS LEARNED)

- Acoustic-phonetic database

  Strengths:
  -- lasting general value for phonetic recognition research
  -- large number of speakers

  Weaknesses:
  -- expensive and time-consuming phonetic transcription
  -- small amount of data

- Naval battle management task domain

  Strengths:
  -- supported development of speaker independent recognition
  -- low perplexity
  -- good demonstration impact

Weaknesses:
-- small language model
-- low perplexity

- "Robust" database

  Weaknesses:
  -- overly ambitious
  -- two-stage collection/computerization
  -- extremely expensive to develop
  -- too large