Will Speech-to-Text Software Work on Audio Recordings from Field Data Collection?

M. Rita Thissen, Sridevi Sattaluri and Carl Fisher

International Field Directors & Technologies Conference

May 20,2008

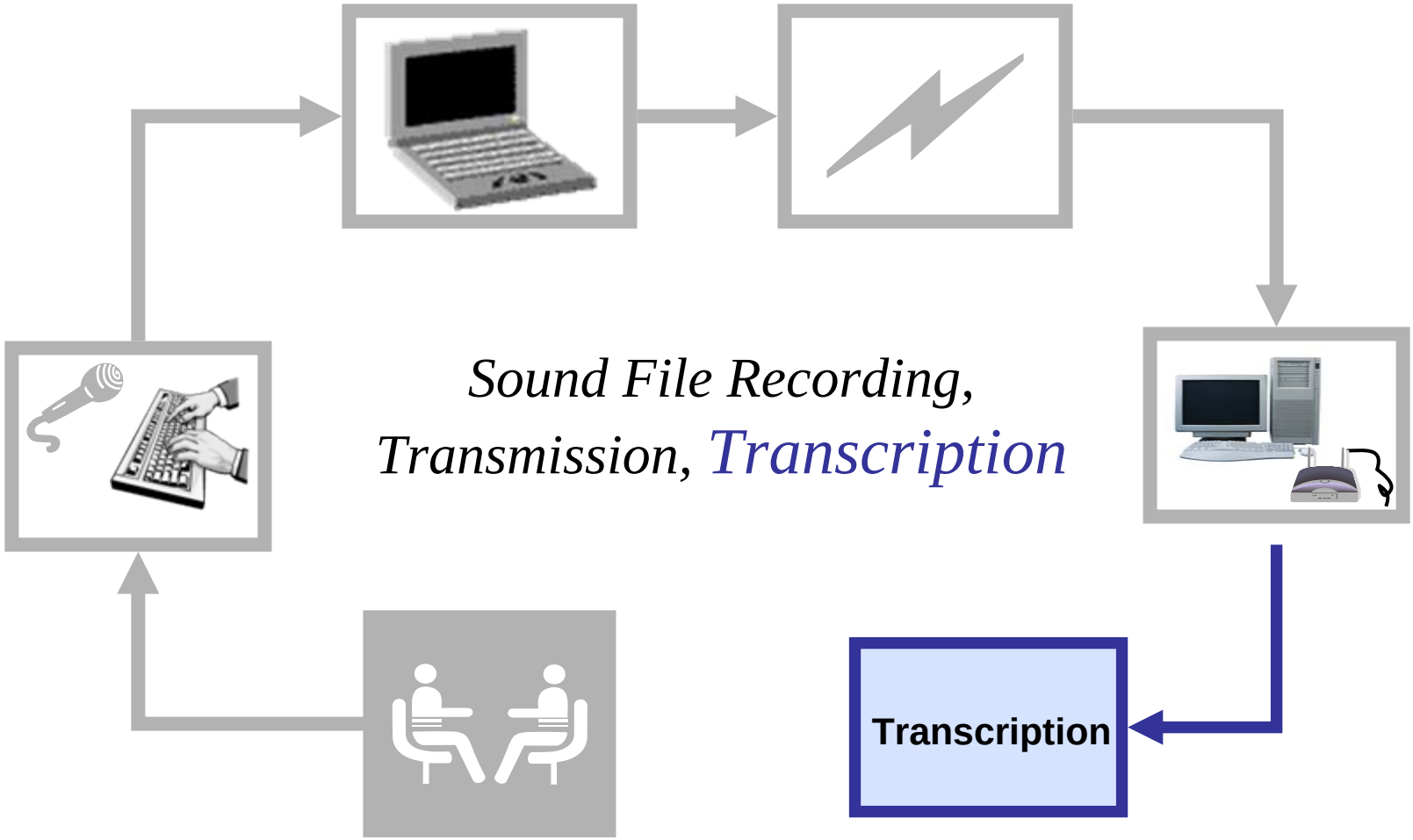
Purpose of This Study

- Investigate the feasibility of using automated transcription software to convert the audio recordings of field interviews into text
- Evaluate the use of ‘Speech to Text’ technologies currently available
- Create automated processes for converting the audio recordings to text
  - To help reduce the costs of the review process
  - To code responses open ended questions

Computer Audio-Recorded Interviewing (CARI)

- Begun by Research Triangle Institute (RTI) in 1999
- Captures audio data for question and response during CAPI interview without interruption
- Provides a cost effective way to monitor the performance of interviewers on the field
- Provides a mechanism for detecting falsification
- Helps evaluate the questionnaire design
- Can be used to capture open-ended responses

Transcription: Wishful Thinking?



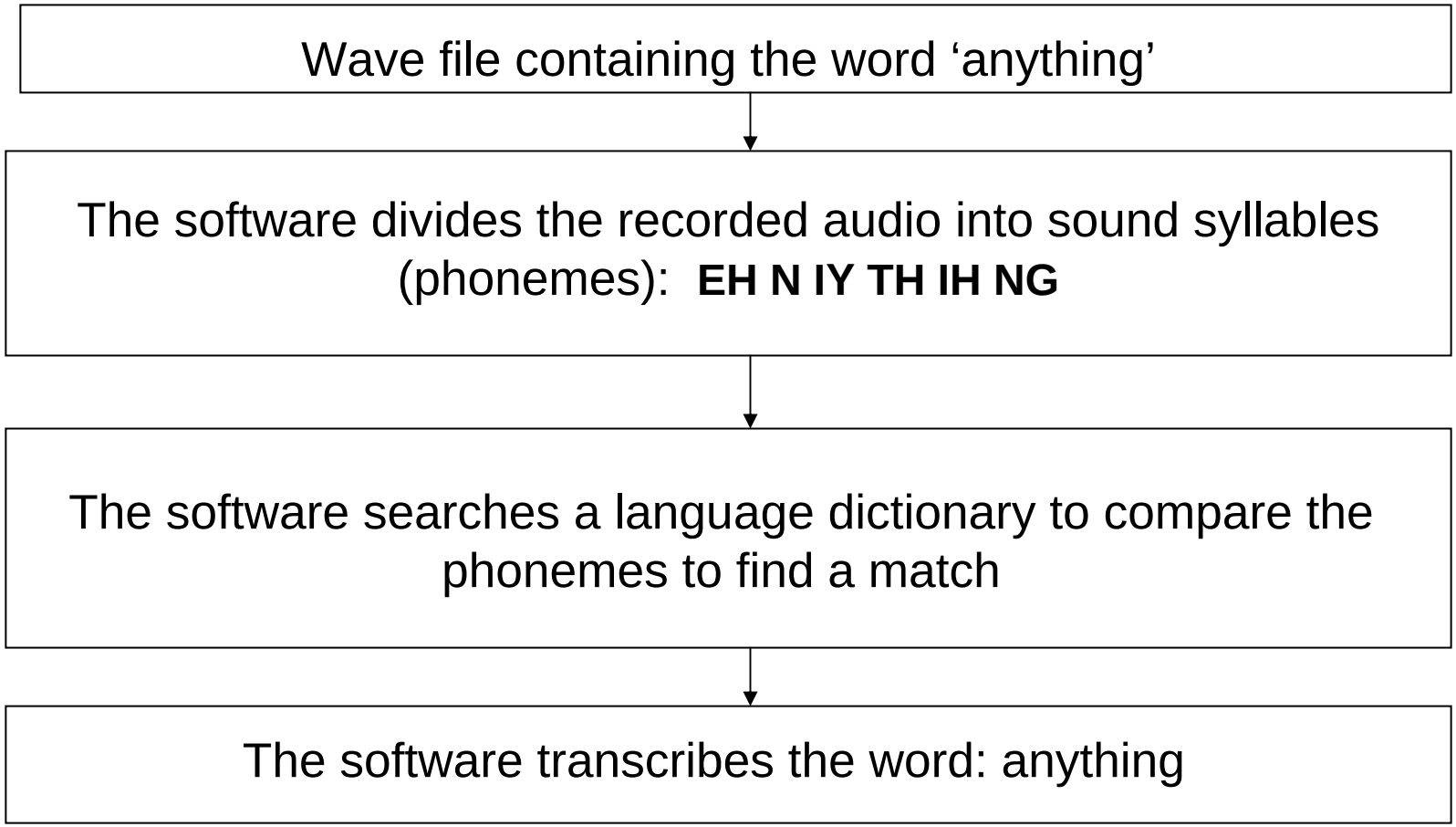
CARI File Characteristics

- Two or more voices in each file
- Many voices in the collection of files
- Background noises from computer or environment
- Uneven loudness of speech
- Variety of accents
- Varying speed of speaking
- Recorded in a specific file format (e.g., wav, 11KHz, mono)
- Large numbers of files (several per interview)

Software Testing Approach

1. Preliminary look
   - Does it handle pre-recorded files or only live audio?
   - Can it handle multiple voices?
   - Can it handle background noise?
   - Does it require special hardware or training?
2. In-depth testing
   - Creation of standard audio files
   - Categorization of voices
   - Transcription
   - Evaluation of accuracy

How Speech-To-Text Works



What Did You Say?

- Multiple phrases may have the same phoneme equivalents: HH OW HH AH R D IH Z IH T OO R EH K UH EY Z B EE CH

  Translation:

  How hard is it to recognize speech?

  How hard is it to wreck a nice beach?

- What are “filled pauses”?

  I don’t, uh, mm, really know.

Software for Converting Recorded Speech To Text

Dragon Naturally Speaking

- Commercial product
- http://www.nuance.com/naturallyspeaking/

Sphinx

- Open-source product developed by Carnegie Mellon University
- http://cmusphinx.sourceforge.net/html/cmusphinx.php
- http://www.speech.cs.cmu.edu/

IBM Nuance Via Voice

- http://www.nuance.com/viavoice/
- Not evaluated in this study (not intended for use with recordings)

Dragon Naturally Speaking: A Quick Exploration

- Commercial product with several editions (we used “Professional”)
- Voice training required and lengthy
- Microphone selection is limited
- Interactive mode is most often used
- Folder-based file processing offers efficient automation
- Background noises (him him him him him)
- Difficulties with multiple voices, changing speeds, changing pitch, changing volume – expects clearly enunciated and uniform speech

Sphinx: A Quick Exploration

- Developed at Carnegie Mellon U, supported by the open-source community
- Speaker-independent
- Hardware tolerant (no specific microphone needed)
- Tolerant of background noise
- Example programs for transcribing live voice and wave files are provided
- Command line interface available for programming
- Supports usage of specific dictionaries and “language models”
- Uses Java Technologies and Sun speech API

Testing Plan

Dragon Naturally Speaking was eliminated from consideration, and we selected Sphinx for in-depth testing.

1. Mock interviewers and respondents are RTI staff.
2. A “Language Model” is built from questionnaire specs.
3. “Standard” (reusable) wave files were collected using Blaise, CARI and a field laptop (IBM Thinkpad).
4. Wave file sampling rate was converted from 11KHz to 16KHz
5. Sphinx software transcribes the wave files.
6. We review the text output for accuracy.
7. We look for ways to improve accuracy and go back to step 5.

How Sphinx Works

- Acoustic model: statistical representation of each possible sound (phoneme) in spoken words
- Dictionary: all the acceptable words in the vocabulary and their phoneme representations
- Language Model: probabilities assigned to each word or sequence of words
- Grammar: all the acceptable words and phrases that a user might say

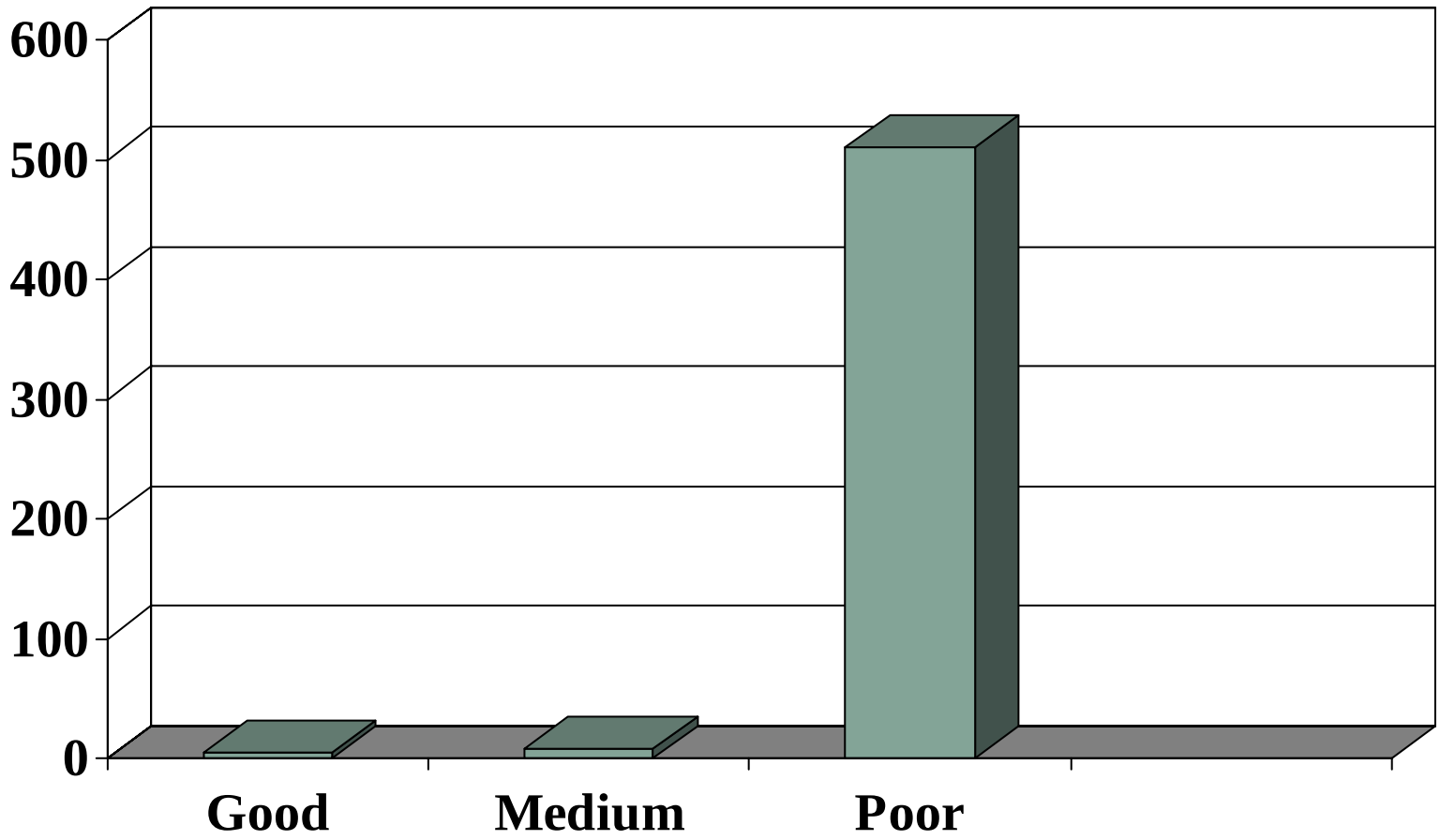
Sphinx Tests: Processing Details

- Mock instrument recorded one sound file per question, including the response
- Sphinx demo program to transcribe .WAV files was adapted to for use in batch mode
- Supporting grammar, dictionary and language model were built from the questionnaire specification
- Sound files were converted to the recording standard expected by Sphinx (16KHz, 16 bit, mono)

Voice Characteristics

Sex:	11 male, 21 female
Age:	6 aged < 40, 15 aged 40-49, 11 aged 50+
Pitch:	5 low, 26 moderate, 1 high
USA:	23 mostly or always lived in USA, 9 did not
English:	24 spoke English at home age 1-14, 8 did not
Loudness:	4 quiet, 25 medium, 3 loud
Speed:	5 slow, 25 moderate, 2 fast
Room:	22 quiet, 9 moderate, 1 loud
Comments:	coughing, air conditioner noise, few interruptions

Results Overall, Trial 1



What is your gender? Female, Male <Response>

- want only gender need
- what here to female
- what and male
- half voice your and
- what your in female
- when united i may and
- would here may
- and what you gender
- voice general male
- what your and or the
- what united female
- what your in a end

- what you interviewer may half
- what year ten there female
- **what is your gender female for male female**
- in male
- **was your gender female**
- in or own and within at an when any what is your gender
- what is your gender mine and
- what is your data female
- question male
- quiet at and female

“Good” Results By Voice Characteristics

Voices varied in the 5 “good” transcriptions

- 2 female, 3 male
- 1 age 40-49, 4 age 50+
- 4 USA, 1 accented (Australian English)
- 2 rapid, 2 medium, 1 slow
- 2 moderate room noise, 3 quiet

Recordings that Transcribed Well

T0002008CARIBlz.Resp_Sex04292008151037.wav

what is your gender female for male female

T0003004CARIBlz.Resp_Sex04302008165439.wav

what is your gender

T0003005CARIBlz.Resp_Sex05012008101434.wav

what is your gender mine and

T0003006CARIBlz.Resp_Sex05012008102214.wav

what is your data female

T0002010CARIBlz.Resp_Sex04302008103013.wav

was your gender female

Is There Any Hope?

Developers' Forum: http://sourceforge.net

Posting: "Standard" CARI file from the mock survey

"To categorize the voice characteristics of the audio recordings, we need to collect some data about me as the interviewer and you as the respondent. For example, my gender is <male/female>. "

Transcription by Nickolay Shmyrev, a SourceForge developer:

"TO CATEGORIZE THE VOICE CHARACTERISTICS OF THE AUDIO RECORDINGS WE NEED TO COLLECT SOME DATA ABOUT ME AS THE INTERVIEWER AND YOU AS THE RESPONDENT FOR EXAMPLE MY GENDER IS MALE "

How did he do it? With a different acoustic model and configuration.

Summary
Future Steps

Conclusion: Not ready for immediate use, but...

- Developer's forum has many tips for improving accuracy – a lot of work is going on internationally.
- Other acoustic models are available (we used American English from the Wall Street Journal).
- Alternate recording formats can be used.
- Transcribing software can be made to handle silences.
- Many options of the software have not been explored.

Acknowledgements

Project team members

Rita Thissen

Sridevi Sattaluri

Carl Fisher

Lillie Barber

Voices of the Research Computing Division at RTI

Sujatha Lakshmikanthan

Support for this work was provided by RTI International Internal Research and Development Funds

Technical support from open-source developers at SourceForge

Questions? Contact Info

If you are interested in automated transcription of audio files and want more information, please contact

Rita Thissen *rthissen@rti.org*

Sridevi Sattaluri *ssattaluri@rti.org*

Carl Fisher *carlf@rti.org*

Slides are available at http://www.rti.org/publications

Thank you for listening!