RTI INTERNATIONAL | turning knowledge into practice

Investigating the Use of Text to Speech Software with a Blaise Instrument

Patricia LeBaron[1], Chuchun Chien[1], Joel Kennet[2], Marty Meyer[1], Dicy Painter[2], Gilbert Rodriguez[1] and Mai Wickelgren[1]

[1] RTI International [2] Substance Abuse and Mental Health Services Administration

Introduction

- In Spring 2009 RTI International and SAMHSA conducted an investigation into the features and quality of various Text to Speech (TTS) software packages in order to determine which, if any, software is best suited for use with the National Survey on Drug Use and Health (NSDUH).

About NSDUH

- Annual face-to-face, household survey.
- Sponsored by the Substance Abuse and Mental Health Services Administration.
- Study is the nation's leading source of information on substance use behaviors and mental health.
- Approximately 67,500 NSDUH interviews are completed annually and data are collected in all 50 states plus the District of Columbia.
- Administered with CAPI and ACASI components.

Research Question

- Has TTS software advanced to the point where NSDUH can transition from using a human voice for the English and Spanish ACASI portions of the survey to using an automated voice created through TTS software?

Why Switch to TTS?

- Human questioner (“voice”) adds time to development process for new questions.
- “Voice” may have limited time and availability for project.
- “Voice” must be retained over several years so that multiple voices are not presented in ACASI when changes are made.
- External factors (relocation, illness, death, others) create risks that may threaten production and costs.
- Upcoming NSDUH redesign may provide opportunity to switch to TTS without worries about potential effects on trend estimates.

Initial Ideas

- 2 methods of using TTS technology

- #1 Real Time
  - The TTS software would create the audio in real-time during the survey.
  - Define pronunciations ahead of time.
  - Load identical versions of the TTS software with the pronunciation definitions on all laptops.
- #2 Recorded
  - Create the audio files centrally using TTS software and load the files onto the field laptops individually.

Why the 2nd Alternative?

- Decided to proceed with option #2.
- Concerns about standardization – unacceptable level of unpredictability with the 1st approach
- Similar level of effort for creating pronunciations
- However, this approach calls for maintaining and storing the wav files needed for the interview from year to year.

Why the 2nd Alternative? Con't

- However, it was decided that the risk of a violation of standardization would outweigh any additional time required to maintain the bank of audio files.

- Therefore, this investigation identified packages that would be suitable for creating customized audio files for storage and dissemination to the field.

Methodology

- Evaluated six TTS software packages
- Acapela, AT&T Natural Voices, Cepstral, Loquendo, NeoSpeech, and Real Speak.
- Focused investigations on female voices offered by the companies.

Methodology, con't

- Created sample audio files using available demo versions of six TTS software packages
- The audio files were created using the text from sample NSDUH questions.
- The wording was quite simple, and we expected the TTS audio of this question to be of a high quality. However, that was not the reality.

Initial Evaluation

- None of the packages provides voices that are free from problems.
  - Robotic
  - Stuttering
- Spanish voices were also problematic
  - Mispronunciations
  - Heavy Spaniard accents
  - An inappropriate tone that may sound rude to respondents
  - Trembling voices

Initial Evaluation, con't

- Would the quality of the audio be suitable for NSDUH?
- Extensive effort required for customizing pronunciations of drug names and other nonstandard words. The English instrument currently uses 3500 .wav files.
- Need to estimate the amount of effort needed to customize pronunciations.

Further Evaluation

- The next steps in our investigation included focusing on just three packages.
- Rated the audio created by the products.
- The English voices were ranked by six evaluators. Each evaluator was asked to rank the TTS voice on each dimension using a scale from 1 to 5. The points on the scale were labeled as follows:
- Numeric Value Label

  1 Poor 2 Fair 3 Good 4 Very Good 5 Excellent

Further Evaluation, con't

- Three Spanish-speaking evaluators scored Spanish wav files

- Each of the three products are:
  - compatible with existing systems,
  - priced competitively,
  - and have potential to be customized to correctly pronounce and emphasize words contained within the NSDUH interview.

Example Audio: Human Voice

- Have you ever, even once, had a drink of any type of alcoholic beverage? Please do not include times when you only had a sip or two from a drink.

- Please look at the tranquilizers shown in Box 3. Have you ever, even once, used Valium or Diazepam that was not prescribed for you or that you took only for the experience or feeling it caused?

English Product Evaluations

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.83	2.67	1.83
Inflection	2.67	3.33	2.33
Tone	3.00	3.17	2.17
‘Humanness’	2.83	3.00	2.50
Pace	3.83	3.00	3.83
Product Average	3.03	3.03	2.53

Spanish Product Evaluations

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.00	1.67	1.67
Inflection	2.00	1.33	1.67
Tone	2.00	1.33	1.67
‘Humanness’	2.00	1.33	1.33
Pace	2.33	1.33	1.33
Product Average	2.07	1.40	1.53

English Product Evaluations: Product #1

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.83	2.67	1.83
Inflection	2.67	3.33	2.33
Tone	3.00	3.17	2.17
‘Humanness’	2.83	3.00	2.50
Pace	3.83	3.00	3.83
Product Average	3.03	3.03	2.53

Spanish Product Evaluations: Product #1

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.00	1.67	1.67
Inflection	2.00	1.33	1.67
Tone	2.00	1.33	1.67
'Humanness'	2.00	1.33	1.33
Pace	2.33	1.33	1.33
Product Average	2.07	1.40	1.53

Example Audio: Product #1

- Have you ever, even once, had a drink of any type of alcoholic beverage? Please do not include times when you only had a sip or two from a drink.

- Please look at the tranquilizers shown in Box 3. Have you ever, even once, used Valium or Diazepam that was not prescribed for you or that you took only for the experience or feeling it caused?

Product Evaluations: Product #1

- The product average for Acapela's English female voice tied NeoSpeech's for the highest ranking. Acapela's Spanish female voice also was ranked the highest among Spanish female voices.

- Acapela's Spanish female voice was described as "trembling."

- Acapela software possesses the capability to edit pronunciations for abbreviations and exceptions. It is batch processing supported and has an adjustable speaking rate and an adjustable voice tone. Acapela includes speech enhancements and control tags.

English Product Evaluations: Product #2

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.83	2.67	1.83
Inflection	2.67	3.33	2.33
Tone	3.00	3.17	2.17
'Humanness'	2.83	3.00	2.50
Pace	3.83	3.00	3.83
Product Average	3.03	3.03	2.53

Spanish Product Evaluations: Product #2

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.00	1.67	1.67
Inflection	2.00	1.33	1.67
Tone	2.00	1.33	1.67
'Humanness'	2.00	1.33	1.33
Pace	2.33	1.33	1.33
Product Average	2.07	1.40	1.53

Example Audio: Product #2

- Have you ever, even once, had a drink of any type of alcoholic beverage? Please do not include times when you only had a sip or two from a drink.

- Please look at the tranquilizers shown in Box 3. Have you ever, even once, used Valium or Diazepam that was not prescribed for you or that you took only for the experience or feeling it caused?

Product Evaluations: Product #2

- NeoSpeech's female Spanish voice was ranked relatively low, with a score of 1.4. This was the lowest ranking of a female Spanish voice.
  - “Too loud or too exclamatory" and
  - “May sound rude to respondents."
- English female voice was described as "assertive."
- NeoSpeech software supports customization of a dictionary so that developers can adjust pronunciations of symbols, abbreviations, and new terms.

English Product Evaluations: Product #3

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.83	2.67	1.83
Inflection	2.67	3.33	2.33
Tone	3.00	3.17	2.17
‘Humanness’	2.83	3.00	2.50
Pace	3.83	3.00	3.83
Product Average	3.03	3.03	2.53

Spanish Product Evaluations: Product #3

	Acapela	NeoSpeech	AT&T Natural Voices
Clarity	2.00	1.67	1.67
Inflection	2.00	1.33	1.67
Tone	2.00	1.33	1.67
'Humanness'	2.00	1.33	1.33
Pace	2.33	1.33	1.33
Product Average	2.07	1.40	1.53

Example Audio: Product #3

- Have you ever, even once, had a drink of any type of alcoholic beverage? Please do not include times when you only had a sip or two from a drink.

- Please look at the tranquilizers shown in Box 3. Have you ever, even once, used Valium or Diazepam that was not prescribed for you or that you took only for the experience or feeling it caused?

Product Evaluations: Product #3

- AT&T's English female voice had low scores on clarity.
- AT&T's Spanish voice was described as having a heavy Spaniard accent and included mispronunciations.
- AT&T Natural Voices includes custom dictionaries that can be used to define pronunciations phonetically.

Product Evaluations: Product #3

- AT&T offered a ‘prescription drug module’ for $1,000. Of the drugs currently used in the four psychotherapeutic modules and in the special drugs modules, 49 out of 99 are an exact match.

- More are partial matches. For example, AT&T's Wizzard Module contains Darvocet, Tylenol, and Codeine, while NSDUH uses Darvocet-N and Tylenol with Codeine.

Customized Testing

- Requested evaluation versions of these 3 software packages.
- Allowed us to customize the pace, pronunciation, and tone for English voices only for Acapela and NeoSpeech.
- An evaluation version of the AT&T Natural Voices product was not available. The demo version had limited capability.
- Created recordings from NSDUH's tranquilizer module and customized the audio files to maximize quality.
- Programming staff tracked the level of effort that was required to achieve the highest possible quality in the end product.

Customization of Software

- Pronunciations of the words were defined either phonetically or using pronunciation symbols within the lexicon/pronunciation editor of the particular application.
- Pronunciations of drug names were particularly laborious to customize, as the default pronunciation of these words differed dramatically from the correct pronunciation.
- It is estimated that each drug name required between ten and twenty minutes to define.

Customization of Software, con't

- Additional time is required to customize other words and pronunciations in the question.
- After defining pronunciations and customizing the results, we rated the resulting audio files in a similar manner.
- The review and rating process also elicited some negative comments about the audio files. Even the customized pronunciation of most drug names presented a problem.

Customization of Software, con't

- Comments about the customized Acapela voice:
  - Difficulty with pronouncing most prescription drug names
  - Problems with inflection at the end of a question
  - A strange accent when pronouncing the word "caused."
  - Voice seemed "laborious" while reading the question

- Comments about the NeoSpeech product:
  - Pronunciations of the drug names were not accurate,
  - In one particular .wav file, the voice sounded as though she talked with a lisp.

Observations

1. Producing files from question text is quick, but the result is a robotic voice that mispronounces many words. Customization is needed.

2. Even after fairly extensive tweaking and customization of pronunciations using phonetic spellings and pronunciation enhancements, the quality of the recordings is sub par.

Observations, con't

3. Customization of simple question text is quite time-consuming.
   - Optimistically, we estimate 1 to 2 hours per question to optimize the pronunciations and cadence of the questions.
   - This estimate does not account for fills or iterative review.

4. With approximately 3,500 English audio files in our questionnaire and 3,500 Spanish audio files, it would take 5 people working full time for more than 1 year to finish the job.

Observations, con't

5. Creating Spanish audio files is challenging.
   - Large level of effort
   - Lack of accurate Spanish pronunciations
   - Limited availability of Spanish speakers who can customize the files
6. The option to purchase Wizzard's prescription drug module might be a time-saver.
   - Only includes English drug names.
   - In cases where the Spanish pronunciation differs from the English, the customization process would be manual.

Conclusions

- Based on observations, TTS technology has not advanced enough for production of NSDUH audio files.
- Very high levels of effort would be needed to customize audio files so that they were of a quality appropriate for use with NSDUH.
- Research is needed to determine what, if any, effects a TTS voice might have on substance-use reporting.
- As a result of the level of effort that would be involved and the quality of the audio, the decision was to not pursue this path.

References

- Couper, M. P. (2005). Technology trends in survey data collection. Social Science Computer Review, 23, 486-501.

- Couper, M. P., Singer, E., & Tourangeau, R. (2004). Does voice matter? An interactive voice response (IVR) experiment. Journal of Official Statistics, 20, 551-570.

Thank You!

Patty LeBaron

plebaron@rti.org