The Speech IVR as a Survey Interviewing Methodology

Jonathan Bloom
Nuance Communications
Draft 9/28/06

Introduction

Speech recognition technology has made great strides in quality over the past decade, and usage has increased in step with this improvement. The technology has many applications, but it has made its most noticeable impact on the telephone, acting as a virtual agent for large businesses and government agencies. Most readers of this chapter have likely called and spoken to an interactive voice response system (IVR) that recognized their speech - or did not recognize it, as the case may be. These speech IVR's have become commonplace, answering calls from individuals when they contact airlines, banks, telephone companies, their county's government, insurance companies, and even prisons[1].

However, in the field of survey interviewing, speech IVR's have not kept apace. Of the larger government surveys, speech IVR's are being used today in surveys for the Current Employment Statistics (CES) program, but only for businesses that do not have access to touchtone. Not many other established surveys are using the technology. This is likely due to several factors, which we will discuss later. But one major reason worth noting up front is that the survey community is still quite busy validating the merits of touchtone IVR's – i.e. phone systems that respond exclusively to telephone keypad input. Until the costs and benefits of this more established technology are clarified, speech IVR's will likely be seen as "around the corner".

Once the research into touchtone IVR's settles down, speech will certainly receive more attention – positive or negative. Survey researchers should be ready with an agenda for vetting the technology and ensuring that it meets certain criteria required for obtaining valid survey data. The primary goals of this chapter are to (1) lay out the inherent strengths and weaknesses of speech technology, (2) point out possible benefits and risks of applying the technology to survey interviewing, and (3) compare it to two closely related survey methodologies – touchtone IVR's and computer-assisted telephone interviews (CATI). Some of the points regarding speech technology in the survey realm will be backed by data, but others are merely the author's estimation because the data do not yet exist.

One final note before we begin: Although researchers are generally quite enthusiastic, the fit between speech technology and survey interviewing is far from given at this point. We should not just be asking whether speech IVR's can be used for the purpose of survey interviewing, but we must also ask *why* we would use it in this way. As Mick Couper points out in this volume, any new technology must be at the service of better data, better experience, and greater efficiency. If we do not feel that speech IVR's hold this promise, then the survey community must look elsewhere for these improvements.

[1] Maricopa County Prison's phone line even offers a spoken option for "self surrender".

A brief introduction to speech recognition

Speech recognition technology involves the mapping of spoken language to text stored in computer memory.[2] The computer's memory includes some number of words or phrases in text form, called the "grammar" or "vocabulary". The goal is to map the incoming acoustic signal from a speaker's voice to the right word or phrase in the grammar.

Let's say a person calls a speech IVR to take a survey and is asked for the state in which she resides. The caller says "California". The speaker's acoustic signal is first picked up by the phone's microphone and transmitted to a computer somewhere in another location. The acoustic signal then goes through "feature extraction" where it is analyzed for certain (quite abstract) qualities, or parameters. Each incoming speech signal varies in its parameter values. The items in the computer's grammar (e.g. "Massachusetts", "Nevada" "Repeat that") also have parameter values associated with them. The values of the words in the grammar were obtained earlier by analyzing potentially thousands of recorded utterances of those words or phrases.[3] The values of the incoming signal ("California") are then compared to the values of items in the grammar. The computer returns an item from the grammar that most closely matches the incoming signal and also returns a "confidence score" for the returned item. As the name suggests, the confidence score indicates the recognizer's confidence that it picked the right match out of the computer's grammar. Hopefully, the system picks "California" out of its grammar, the confidence score is high, and thus the system officially recognizes "California". The grammar changes with each question, looking quite different for a yes/no question than for a request for a zip code.

Many variables can compromise the confidence of the recognition, and they are more or less the same variables that can compromise human hearing. For example, the speaker may be located in a noisy environment such as a car or a house full of children and pets.

Also, people do not speak fluently. Speech production is rife with false starts, repairs, and filled pauses such as 'um' and 'uh'. There is some evidence that people are more fluent when faced with a speech recognition system (Oviatt, 1993), but there is little that a person can do to avoid sneezing, coughing, etc. Because speech IVR's utilize telephones and phone networks to carry the speech signal to the recognizer, the signal may also be compromised by a bad connection. Speech recognition systems are trained on many different accents and dialects, but very strong accents may also be met with degraded accuracy.[4]

[2] This is often confused with its counterpart "text to speech" (TTS), in which text in the computer's memory is converted into speech by way of a synthetic voice. We will also briefly discuss TTS.

[3] This is quite a lot of overhead to consider for a small survey interview. However, most speech recognition software on the market today comes with grammars that have default "language models" built in. Accuracy will be only "good" at first, but will improve as more data – data specific to the survey's calling population – begin to come in.

[4] Unlike speech recognition used for dictation software, speech IVR's also cannot take advantage of linguistic context to help identify a word. Dictation software can look at preceding and subsequent words to narrow down the possible matches for a target word.

The confidence can also be compromised by the size of the grammar and the confusability of the items in the grammar. Accuracy on yes/no questions is usually quite high because there are only two words (and perhaps a handful of synonyms) that the system can recognize, and also because the words do not sounds like one another. Accuracy is comparatively lower when the grammar includes, for example, all of the train stations in the United States served by Amtrak®. There are hundreds of items in the grammar and some of them are highly confusable (e.g. Newark Penn Station and New York Penn Station).

Poor recognition accuracy is obviously one of the biggest challenges facing speech IVR's as a legitimate means of survey interviewing. One of the major benefits of the telephone as a medium for interviewing is supposed to be that people can easily respond to survey questions from their homes or on the road. The noise of an active home a busy street, or a bad connection should not stand in the way much more than it does for human interviewers on recognition accuracy measures if speech recognition is to be a sufficient method of survey data collection. Studies suggest that human listeners score on an order of magnitude better on recognition accuracy measures than do speech recognition systems (Lippmann, 1997 as cited in Moore, 2003; Meyer et al., 2006). Meyer et al. (2006) compared the recognition accuracy of computers and humans using a database of three-syllable nonsense words that varied in speaking rate, volume, intonation, and dialect. Computer and human participants were required to recognize the middle phoneme of each nonsense word. They found an average accuracy rating of 99.4% for human listeners and 74.0% for the speech recognition system. Other studies also conclude that, whereas speech recognition technology varies quite a bit in accuracy depending on variables like noise levels, human recognition is more robust, staying consistently high even in the face of the aforementioned hurdles (Deshmukh et al., 1996).

This problem may seem daunting, but as we will see in the next section, speech technology can greatly improve its chances by exploiting many of the conversational strategies that human addressees use when faced with uncertainty about a speaker's utterance.

Speech recognition applied to IVR's

So far, we have discussed speech IVR's at the level of a single respondent input. A speech IVR is a dialog, and so it must not only be the listener, but must also take turns being the speaker. It must be capable of speech output. Output for speech IVR's (and touchtone IVR's) comes in two forms: recorded human speech or computer-generated speech, also known as text-to-speech (TTS). The effects of using TTS versus human recordings on survey data are still being debated (see Couper, Singer & Tourangeau, 2004). But it should be noted that TTS improves at an impressive rate, becoming more understandable with each new release. Therefore, the results of the studies today could require updating fairly soon. TTS is reaching a turning point, after which the distinction between it and human recorded speech may become undetectable. Recently, developers of TTS started using a new strategy which strings together phonemes of actual human voices. This new "concatenative" approach to TTS has made a marked improvement in the quality of TTS when compared to traditional “formant” methods, in which all TTS audio is generated by the computer.

The speech IVR cannot just recognize speech and generate speech. Speech IVR’s try to simulate human conversations, and human conversations are rule-governed activities (e.g. Clark, 1996; Sacks et al., 1974). In the case of a survey interview, the rules may be quite simple when the respondent can hear the question clearly and the respondent and the interviewer understand the question the same way, or at least both parties assume they do. Take for example this portion of an interview from a survey regarding tobacco use and opinions about tobacco use (Suessbrick, 2005):

I: About how long has it been since you last smoked cigarettes every day.

R: Um: s: uh since . I think February of ninety-two: I quit so: its like nine: . two three four five six seven eight nine, seven years.

I: Alright.

The dialog goes quite smoothly. The interviewer poses a question and the respondent, after some calculations, answers “seven”. The interviewer then provides a backchannel, “Alright,” to let the respondent know that their answer has been heard and recorded.

If a speech IVR only needed to play out a question and recognize (a more constrained version of) the above answer, the job of handling a survey interview would be quite simple. However, consider the following question and answer that takes place further along in the same interview.

I: About how long has it been since you COMPLETELY stopped smoking.

R: . Since um nineteen ninety-two.

I: Okay, so how many . years would that be

R: That's um: . two three four five six seven eight nine, that's seven years, wait didn't you ask me that same question before?

I: No we asked you uh . this is w- we asked you how long you had smoked but now we're asking you about how long its been since you completely STOPPED smoking.

R: Okay.

I: Do you think that was *for*

R: *No no* I- it was another question where I answered seven years?

I: Oh you think maybe you misunderstood that ques*tion*?

R: *Yeah* uh I don't want to uh.

I: Well it's about how long has its been since you last . smoked cigarettes every day

R: Right, oh okay, got it

I: Okay so that was *uh*

R: *Yeah* that was still seven *years*

I: *Alright.*

R: Mm-hm.

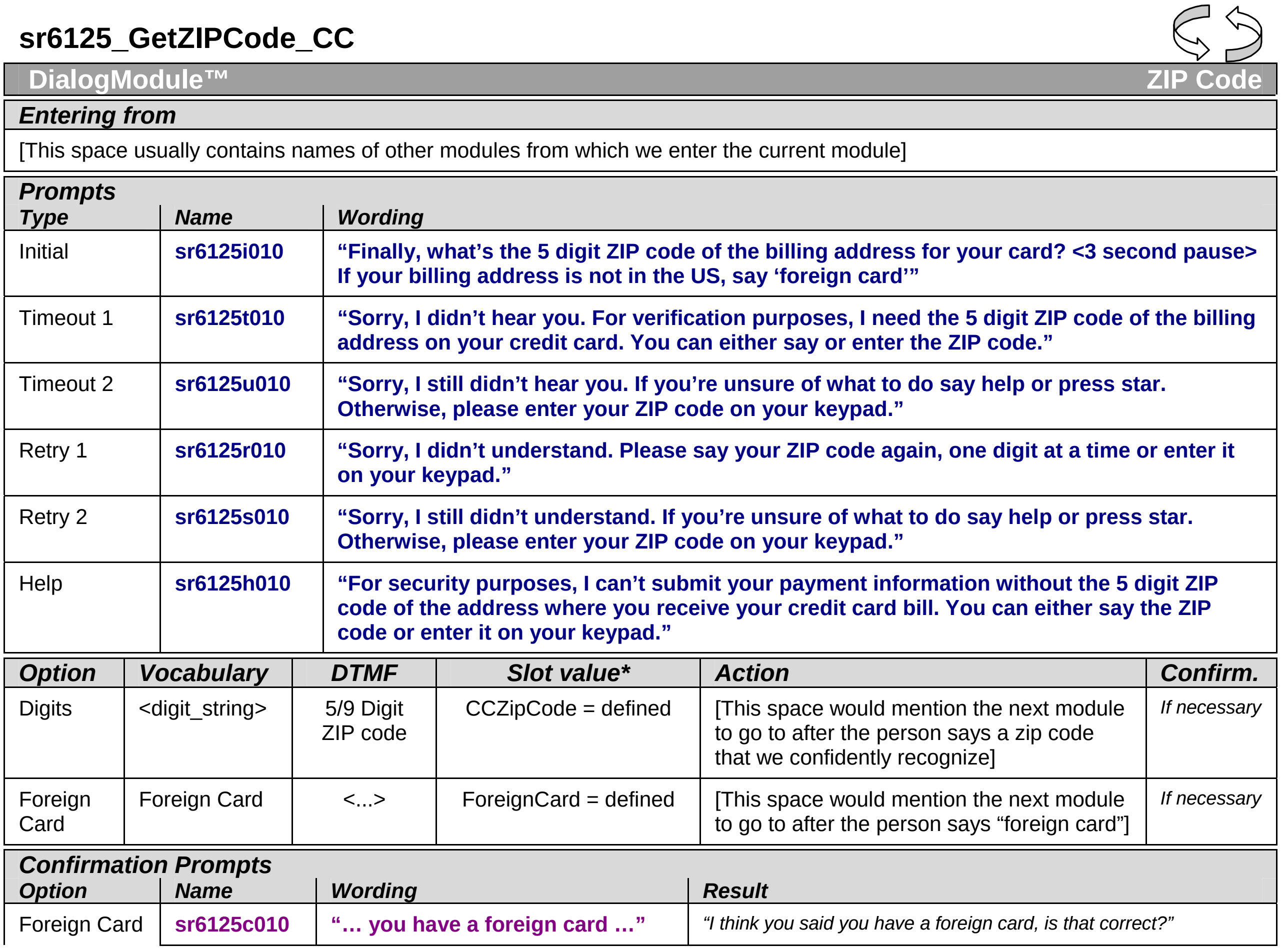
I: Now again, how about how long has it been since you completely stopped smoking cigarettes . sort of, R: Yeah. Seven years.

I: Alright.

R: And a couple of months.

During this interaction, it becomes clear that the respondent misunderstood an earlier question and the respondent and interviewer must now return to answer it again. The respondent may not have heard the question properly the first time, or heard it but did not understand it. Conceptual misalignments between interviewer and respondent do happen, are especially hard to resolve when traditional approaches to standardization (as defined by Fowler & Mangione, 1990) are followed (e.g., Conrad & Schober, 1999, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker, 2004). In one instance, Suessbrick, Schober, and Conrad (2000) found conceptual misalignments occurring almost as often as alignment.

Given that respondents may not hear a question or may not understand a question, or may think they understand a question when they do not, the rules that govern turn-taking in speech IVR's must be flexible. The speech IVR's of today cannot hope to handle the dialog above, but most of them do have error handling in place that can address many common problems. Figure 1 shows the basic inner workings for one conversational turn in a speech IVR that handles train travel. The prompts that are played out to the caller are bolded.



sr6125_GetZIPCode_CC

DialogModule™ | ZIP Code

Entering from

[This space usually contains names of other modules from which we enter the current module]

Prompts

Type	Name	Wording
Initial	**sr6125i010**	**"Finally, what's the 5 digit ZIP code of the billing address for your card? <3 second pause> If your billing address is not in the US, say 'foreign card'"**
Timeout 1	**sr6125t010**	**"Sorry, I didn't hear you. For verification purposes, I need the 5 digit ZIP code of the billing address on your credit card. You can either say or enter the ZIP code."**
Timeout 2	**sr6125u010**	**"Sorry, I still didn't hear you. If you're unsure of what to do say help or press star. Otherwise, please enter your ZIP code on your keypad."**
Retry 1	**sr6125r010**	**"Sorry, I didn't understand. Please say your ZIP code again, one digit at a time or enter it on your keypad."**
Retry 2	**sr6125s010**	**"Sorry, I still didn't understand. If you're unsure of what to do say help or press star. Otherwise, please enter your ZIP code on your keypad."**
Help	**sr6125h010**	**"For security purposes, I can't submit your payment information without the 5 digit ZIP code of the address where you receive your credit card bill. You can either say the ZIP code or enter it on your keypad."**

Option	Vocabulary	DTMF	Slot value*	Action	Confirm.
Digits	<digit_string>	5/9 Digit ZIP code	CCZipCode = defined	[This space would mention the next module to go to after the person says a zip code that we confidently recognize]	*If necessary*
Foreign Card	Foreign Card	<...>	ForeignCard = defined	[This space would mention the next module to go to after the person says "foreign card"]	*If necessary*

Confirmation Prompts

Option	Name	Wording	Result
Foreign Card	**sr6125c010**	**"… you have a foreign card …"**	*"I think you said you have a foreign card, is that correct?"*

Digits	Default confirmation, as handled by DialogModule™	*"I think you said zero six eight five three, is that correct?"*

Figure 1
Amtrak module handling zip code collection. Spoken prompts are bolded.

This module handles the play-out of one "initial" prompt that requests information from the caller. In this case it is "Finally, what's the 5 digit ZIP code of the billing address for your card? <3 second pause> If your billing address is not in the US, say 'foreign card'." In a survey interview, this would be where the actual survey question is asked, for example, "About how long has it been since you *completely* stopped smoking?"

The respondent may do many things in response to the initial prompt. One possibility is that they do nothing. They do not speak. If the respondent does not know how to answer the question, this is a distinct possibility. In such cases, speech IVR's have a fallback called a "timeout 1" prompt which asks the caller the question again. One commonly used strategy in timeout prompts is to elaborate on the original question, trying to predict why the caller may not have answered. However, if this were a traditional survey interview in which clarification of a question is not permitted, the timeout prompt may need to simply repeat the initial prompt. Survey interview designers need to consider how to take advantage of these timeout prompts. The respondent may understand the question but remain silent because they don't know how to interact with the system. One possibility is that the timeout prompts do not change the question wording, but elaborate on how the respondent can be more successful interacting with the speech IVR. For example, in the module shown above, the system mentions touchtone input as an alternative modality to speech (it says "please say *or enter*"). In doing this, interviewers may be able to increase response rates while keeping question wording the same.

If the caller does not respond again, the module includes a "timeout 2" prompt, often asking the initial question again, but also mentioning either the agent option or a help option which we will discuss in a moment. If the person does nothing a third time, most speech IVR's transfer to a call center agent. If this is not a possibility with survey interviews, the system could tell the person to call back when they are ready or able to respond, or simply keep reusing timeouts 1 and 2 ad infinitum.

Even if the caller does respond, their response may be recognized with "low confidence" for some or all of the reasons mentioned earlier (noisy environment, thick accent, etc.). If the system does not recognize the caller, most speech IVR's come with a built-in "retry" prompt that often looks similar to a timeout prompt. Often the only difference is that the timeout prompts starts with "Sorry I didn't hear you" and a retry prompt starts with "Sorry I didn't understand." Like timeouts, there are also "retry 2" prompts in case the person is understood with low confidence a second time.

There is also a context-sensitive help prompt included in this module, which is played out if the caller explicitly asks for help. The help prompt elaborates on the initial prompt. The retries and timeouts do this as well, but the help prompt is supposed to be even more detailed, the assumption being that if the caller explicitly asked for help, then they are willing to listen to a longer explanation.

Few people interacting with speech IVR's ever explicitly ask for help and so these prompts are rarely heard by callers. That few people explicitly ask for help in speech

IVR's aligns with findings in other types of discourse that suggest people rarely ask for help (Graesser & McMahen, 1993). Nonetheless, for those people who do use it, a help prompt cannot just be a repeat of the initial prompt. For survey interviews, this can be a challenge depending on the survey designer's definition of standardization. As mentioned earlier, Fowler and Mangione (1990) established guidelines of standardization that are followed by many organizations. These guidelines are designed to minimize error by keeping question wording invariable. If the respondent is having trouble answering, the interviewer is only allowed to guide respondents with "neutral probes" like "Whatever it means to you". A speech IVR could be made to respond in such a manner if a respondent asked for help.

But some feel that standardization should not be applied to wording, but to meaning, so that the question is understood the same way by the interviewer and all respondents (e.g. Suchman & Jordan, 1990). More open-ended collaboration of question meaning during survey interviews can lead to higher response accuracy (Conrad & Schober, 2000). If one takes a more collaborative approach, then the help prompt in a speech IVR - as well as timeout and retry prompts - can be put to better use. It may be able to provide more information that answers many of the respondents' questions. Of course, there is evidence that allowing for elaboration on question meaning can make (simulated) speech IVR interviews longer (Bloom, 1999) which may be a cost issue. Also, providing a help prompt is not the same as real-time human-human collaboration. Speech IVR designers must figure out in advance what elements of a survey question might confuse respondents, and put that information in the help prompts. But there is no guarantee that the designers will predict accurately.

In addition to initial, timeout, retry, and help prompts, these modules also include "confirmation prompts". In certain cases, the caller may be recognized by the system with "medium confidence". The confidence was not high enough to simply move on to the next question, but also not low enough to retry. Like human addressees in such situations, the system responds by saying something like "I think you said the zip code was oh six eight five three, is that right?"[5] These confirmation prompts are used pretty commonly, except for yes/no questions, because that can lead to serious confusion ("I think you said 'no.' Is that right?" How does one respond to that if the answer is "no" and then how does the speech IVR interpret that "no" answer?). Confirmation prompts are unique to speech IVR's; they have no equivalent in touchtone IVR's.

One could look at such confirmation prompts in the context of a survey interview and think of them as the equivalent of Fowler and Mangione's "neutral probes". However, whether neutral probes are actually neutral is up for debate. There is evidence from human-human interviews that such probes could be used to lead respondents to a specific answer (Schober & Conrad, 2002). Subtle cues in the interviewer's presentation of the probe can be picked up by respondents and interpreted as hints that the answer being provided is incorrect. In the context of a speech IVR, we cannot be sure if confirmation prompts would be interpreted as leading the respondent one way or another. People may assume computers do not understand or employ such conversational

[5] Designers figure out the medium confidence thresholds by trial and error. They set the distinction between high, medium, and low confidence somewhat arbitrarily at first, and then listen to caller responses to the confirmation prompts. They then set the thresholds so that approximately 50% of the responses to confirmations are 'yes' and 50% are 'no'.

knowledge. If that turns out to be the case, then there is no concern (although assuming less capability on behalf of the IVR system may have its downsides too, e.g. users feeling the need to segment each word when speaking). However, we need to be mindful of such possibilities when considering speech IVR's for the purpose of survey interviewing.

Finally, in speech IVR modules there are usually one or more "global commands" active. Global commands are commands that are available throughout the interaction. For example, "repeat that" is often available in all conversational turns, regardless of context. Where the speech IVR could use retry prompts when confidence was low, people speaking to speech IVR's can use repeat that when *their* confidence is low. "Repeat that" is a powerful resource, given the serial and ephemeral nature of speech (as opposed to parallel and persistent interfaces like the web).

So far, we have taken a high-level tour of speech IVR's. We have seen that speech IVR's are comprised of speech recognition input, TTS or human recorded output, both combined with the aforementioned conversational strategies. In the process we have discussed many strengths and risks of this technology when considering it as a method of gathering survey data. Most of these risks and benefits were inherent in the technology. In the next section, we will take a more comparative approach, itemizing the pros and cons of speech IVR's when compared to touchtone IVR's and to CATI.

Speech IVR's compared to CATI

A large body of research exists comparing surveys administered using touchtone IVR's to those administered using CATI. Far less research has been done to compare speech IVR's to CATI. However, when comparing speech IVR's to CATI, the benefits of speech IVR's may overlap with those of touchtone IVR's, given that both are telephony-based self-administered methods of data collection.

So far, the data suggests that touchtone IVR's do offer some benefits over CATI. Most importantly, interviews using touchtone IVR's cost less than CATI (Miller-Steiger, 2006). Although speech IVR's do cost more to build and maintain than touchtone IVR's (because grammars need to be written and improved over time), the cost is still less than hiring and training survey interviewers (Clayton & Winter, 1992).

In addition, respondents are more likely to provide socially undesirable responses with a touchtone IVR than they are with human interviewers, for example when answering questions about tobacco use (Currivan, Nyman, Turner, & Biener, 2004) and sexual activity (Tourangeau & Smith, 1998; Villarroel et al., 2006). At first blush, it appears that "turning down" the social presence of the interviewer makes respondents more willing to admit to socially undesirable behaviors. But the relation between social presence and socially desirable answers may be less a matter of degree and more a binary distinction – "human" versus "not human". Couper, Singer & Tourangeau (2004) found that touchtone IVR's lead to more socially undesirable admissions than using human interviewers. But no difference was found between automated systems that varied in level of "humanity". More specifically, they found no difference in socially undesirable responding when employing IVR's using recorded human speech versus computer-generated speech (also known as text-to-speech, or TTS). This suggests that if people know the "speaker" is not a live human then the similarity of its voice to that of a live human voice does not matter.

Clifford Nass' lab at Stanford used a speech IVR to conduct a similar study, although the human interviewer condition was removed so that all a computer produced all output (Nass, Robles, Bienenstock, Treinen, & Heenan, 2003). They did find an effect of the IVR's "humanity", but they found that disclosures *dropped* when they used TTS versus recorded human speech. It is hard to tell what accounts for the different results given that there are many differences in the methodologies of the two studies. However, one interesting possibility is that the use of speech recognition in the latter study - as opposed to touchtone - impacted the direction of the effect. Is it possible that once we are confined to just comparing automated systems (and no longer comparing them to human interviewers) that what counts is not level of the output's "humanity", but rather the internal *consistency* of humanity across the input and output modalities? Perhaps TTS aligns better with touchtone input and human recorded speech aligns better with speech recognition. When the inputs and outputs of the technology are paired in terms of their proximity to human conversation, respondents may be more willing to open up to the system.

The point is, touchtone IVR's seem to increase respondent willingness to give answers that are socially undesirable, and hopefully, this benefit can be generalized to speech IVR's as well. The question arises whether the addition of speech input (as opposed to touchtone input) might lessen the effect, given that respondents are interacting in a way that is more like what they do with a human interviewer. Because respondents are engaged in a simulated conversation, will this turn on social presence? In addition, people may be less likely to divulge sensitive information with a speech IVR because they are vocalizing. If anyone else is nearby, the respondent may feel uncomfortable speaking their answers. We have found that touchtone entry in our speech IVR's (we usually offer both) is highest when callers are asked for sensitive information, like credit card numbers or account passwords.

Another benefit of speech IVR's over human interviewers (touchtone IVR's share this benefit as well) is the added control the survey designers have over standardization. With speech IVR's, all participants hear the exact same questions presented the exact same way. The variation in presentation across respondents is removed. In addition, as we have already discussed, survey designers can control the amount of elaboration that the system provides when misunderstandings occur. A survey designer can follow the guidance of Fowler and Mangione, or take a more "collaborative" approach (e.g. Schober & Conrad, 1997; 2002) and try to clarify question meaning during fallback prompts. The system can offer whatever amount of elaboration the designers are comfortable providing – up to the limits of what the dialog design can handle This last point is important. If a survey designer wished to create a survey interview that was highly interactive, CATI may be preferred over speech IVR's. The vast majority of speech IVR's are "system-initiated" meaning that the system guides the direction of the conversation. For example, the first words spoken in the conversation are uttered by the speech IVR, not the respondent. From there, the IVR sets the agenda, leaving no room for the respondent to answer questions out of order, or ask their own questions (e.g. "How long will this take?"). However, it is possible, in principle, to design a speech IVR with "mixed initiative" (system and user/respondent control the dialogue); but this would require substantially more advanced dialogue management than what is used in current systems.

(For a discussion of the costs and benefits of mixed initiative in interviews, see Schober, Conrad & Fricker, 2004).

Speech IVR's also have problems when compared to CATI. Most troublesome is the lower accuracy of speech IVR's that we have already mentioned. If a survey population includes many people who will have heavy accents, for example, then speech IVR's are risky.

But speech technology need not be discarded outright for this reason. Accuracy issues vary depending on context. Speech IVR's do quite well with recognition of yes/no responses and digit collection. Therefore, speech IVR's can still be put to good use with specific kinds of data collection. For example, the Current Employment Statistics program (CES) conducted by the US Bureau of Labor Statistics collects employment data from thousands of nonagricultural businesses each month. The nature of the data being collected is relatively simple from a speech recognition standpoint. It is comprised predominantly of digits, which is one of current speech recognition technology's stronger suits. [6]

Also, people learn to interact with speech systems upon repeated use. If a survey is going to be targeting a specific sample longitudinally, then speech IVR's become more viable. Again, the CES stands out as an appropriate survey for speech technology. The same companies respond to the same questions over time. Observing speech recognition data on the CES in 1989, Clayton and Winter (1992) found that error due to recognition problems decreased after a respondent's first month of usage. Although recognition accuracy is a major concern when compared to other modalities, the problem will not be as severe if the technology is applied to the right type of survey, i.e. one collects recognizable data such as "yes"/"no" and numeric responses from the native- and clearly speaking respondents who use the system on a recurring basis.

Aside from accuracy, another likely shortcoming is actually found in research with touchtone IVR's. They show higher breakoff rates than CATI (Miller-Steiger, 2006). This may be because an automated system puts little if any social pressure on a respondent to continue. Apparently, the same "low social presence" that allows for more truthful responses also gives respondents more freedom to hang up. Therefore, touchtone IVR's are preferable for surveys that are shorter in length and include shorter scales (Dillman et al., 2001 as cited in Miller-Steiger, 2006). These limitations are non-trivial, especially in light of the fact that telephone surveys of any kind, CATI or otherwise, lend themselves to shorter scales and surveys (Dillman, 2002). With automated systems, they may need to be even shorter.

Clayton and Winter (1992) again found heartening results when speech was used for the CES. Looking at more than 1000 respondents, they found over the course of two years (1989 and 1990) that response rates for speech, touchtone and CATI were all comparable, with mail lagging far behind (the authors do not report statistical significance). The response rates for speech, touchtone and CATI were all higher than 90%. Speech never dropped below 85% in any given month. This is surprising, given the state of the art at the time of this research. They were using dated TTS technology, and

[6] Also, there is no reason that survey designers must choose between speech *or* touchtone. The two modalities are blended quite successfully in today's commercial IVR's (e.g. "If you're calling about a missed appointment, say 'yes' or press 1 Otherwise, say 'no' or press 2.").

could only allow for digits spoken one at a time – for example “four one two” as opposed to “four hundred and twelve.”

Even if response rates are comparable to those of CATI (and I am somewhat skeptical this will happen any time soon), respondents will most likely prefer talking to a human, at least about non-sensitive topics. In the design of commercial systems, we often hear customers mention that they appreciate speech IVR’s over humans because speech IVR’s do not mumble due to exhaustion or “cop an attitude”. But these individuals are the exception rather than the rule. Anecdotally, we hear far more often that these systems are cold and unfeeling (and that they “pretend” to be warm).

Speech IVR’s compared to touchtone IVR’s

Speech IVR’s offer many benefits over touchtone. For one, people seem to prefer speech over touchtone. At Nuance, we reviewed 25 applications, including 289 individual speech modules, and 1.6 million caller responses. In the cases where prompting explicitly mentioned the availability of both modalities, approximately 70% of callers still used speech on the first try. This percentage did not decline on retries and timeouts that mentioned both modalities.

When Clayton and Winter (1992) asked respondents for their preference between the speech version and the touchtone version of the CES, 60% preferred speech over touchtone, while 32% preferred touchtone, and 10% did not respond. Again, those numbers are interesting considering the dated TTS and single digit input. Their respondents also experienced the speech IVR as shorter, even though the calls were on average 20 seconds longer due to additional instructions.

Another benefit of speech is that it facilitates the input of certain types of data. The CES involves numeric input and yes/no questions, so touchtone and speech are both viable options . What if participants had to provide an address? Entering addresses using touchtone is quite labor intensive. The same can be said for names, dates, and the selection of items from long lists (e.g. names of countries). If a survey includes data of this sort, then speech becomes a much more compelling interface. Borrowing an example from the commercial sector, Amtrak used a touchtone IVR up until 2001. When callers had to specify the arrival and departure stations, they were required to enter the first three letters of both the arrival and departure city names. So if the caller was traveling to Boston South Station, they would need to enter 2-6-7. This requires them to hunt for the numbers on the phone’s keypad that correspond to the three letters. That alone makes the interface more difficult than speech. But also consider that the caller still needs to identify which station in Boston selecting from a menu (there are three stations served by Amtrak in the Boston area). Since switching to speech, the Amtrak IVR now asks for the station name instead of the city, for example “Boston South Station”. It would not have made sense to ask for the station instead of the city name with touchtone, because the first three letters are not enough to disambiguate between the three (Boston South Station, Boston North Station, and Boston Back Bay Station). By using speech, we have turned two steps into one (“city + station” becomes “station”), and made that single step less labor intensive (removed hunting for keys on the keypad).

Because humans are designed to communicate via speech (Pinker & Bloom, 1990; Pinker, 1994), the prompts in a speech IVR can also be shorter than those in a touchtone IVR. No explicit directions are required. For example, with a speech IVR, a yes/no question like “In the past year have you seen a medical doctor” can be asked without any extra wording (the question comes from the Current Population Survey). With a touchtone IVR, additional verbiage is required to explain the options: “In the past year have you seen a medical doctor? Press 1 for ‘yes’ or 2 for ‘no.’” Of course, the caveat here is that humans are not necessarily designed to communicate using speech *to interact with a computer*. With computers, common ground can be harder to establish (see Brennan, 1991; 1998). What words does the system know? What does it remember from the last conversational turn? From the last call? Because of this, directions may be required when the interaction runs into problems, i.e. in timeout and retry prompts.

Speech IVR’s also offer better ergonomics. When a respondent is using a phone that has the touchtone keypad on the handset, they do not need to take the phone away from their heads in order to respond (Mingay, 2000). This is the case for some landline phones as well. For longer surveys, differences like this could have a considerable impact on call duration and also on the respondent’s opinion of the survey.

There is no “confidence level” associated with touchtone input. It is an “all or nothing” modality, where a selection was made or it was not. The certainty of the touchtone IVR could be considered a benefit, but there are negative aspects associated with that certainty. Groves (2005) suggested that the best data that survey researchers could hope for would, among other things, include informative paradata. With speech IVR’s, the messier input can also be seen as rich paradata that can be logged and analyzed. We can hear the respondent’s level of anger or disinterest when they answer a question. This information could also be used to train the recognizer, so that if it later picks up “anger” in a response, it could digress from the survey path and assure the caller that, for example, the survey is almost complete. With touchtone entry, there is no way to tell how angrily the respondent pressed a key.

This is a double-edged sword. The rich paradata provided by speech IVR’s is also their primary problem. Just like the comparison between speech IVR’s and CATI, response accuracy is again an issue when compared to touchtone IVR’s. Assuming that the respondent understands the question properly and assuming that she clearly enunciates her answer in a quiet environment, the recognizer may recognize the right response with high confidence, but it might also correctly recognize the right response with low confidence, or incorrectly recognize the wrong response with high confidence. And we cannot even assume that the respondent’s speech and environment are ideal. Some respondents may respond eloquently, but with a thick accent.[7] The caller may have side conversations (e.g. “Honey, did we buy any furniture in the past year?”). They may be calling from a city apartment with open windows on a major truck route, or they may simply sniff. All of these possibilities add up to longer calls and the risk of data error. Speech recognition technology is improving constantly. Recognizers are getting better at distinguishing signal from noise and accurately interpreting the signal, but this does not mean they are close to reaching the accuracy of touchtone.

[7] Respondents who speak English as their second or third language may also prefer touchtone, possibly even over a human interviewer, because touchtone only makes demands on their English comprehension abilities, not on their production abilities (assuming for a moment that this is a survey in an English-speaking country).

Also, there are language limitations. Speech recognizers exist for many different languages at the time of this writing (e.g. U.S. English U.K. English, Australian English, Spanish, French, German, Portuguese, Hebrew, Japanese, Mandarin Chinese, etc.), but the list is obviously not exhaustive.

Another smaller downside to speech IVR's is the memory overhead required during menu contexts. The caller must listen to the options in the menu and try to rehearse the wording of the most appropriate options while they listen for potentially better options. With touchtone, the caller can externalize their memory by simply placing a finger on a key while listening to the rest of the options.

Finally, there is the question of social presence. At this point it is hard to say what effect speech IVR's will have on social presence when compared to touchtone IVR's and CATI. Speech IVR's could strike a happy medium between CATI and touchtone Speech IVR's more closely simulate a human conversation than do touchtone IVR's, and because of this, social presence theoretically increases. The effect may be that speech IVR's offer the best of both worlds, keeping truthful responses to sensitive questions high due to its automation, while keeping respondent breakoffs low because of its similarity to human conversation. This is, of course, the hope. Medium social presence could also backfire, leading to the worst of both worlds. Speech IVR's could lead to a drop in truthful responses to sensitive questions because they simulate a human conversation and unacceptable breakoff rates because of its automation. These are empirical questions waiting to be answered.

Conclusion

Speech technology comes with many potential benefits and shortcomings. Based on those discussed in this chapter, we can start to identify the kind of situations in which speech IVR's would make sense as an alternative to CATI or touchtone IVR's. I envision the following optimal scenario for speech:

> *A researcher for a pharmaceutical company is curious about the regularity with which people use their products. She wants to get weekly information over a period of one year. She decides to set up a longitudinal survey including a sample of several hundred individuals. Ninety percent speak English as their first language. The researcher has not been given the budget to hire interviewers for an extended period of time, but is concerned about the low response rates of mailed-out questionnaires. The data are sensitive, so the designer feels it would probably be best to have the respondents interact with an automated system. The company has given its blessing for her to work with the IT department and use its phone network. She would consider a web survey, but the population being surveyed includes many busy individuals who do not have time to answer these questions while at work or home, and would prefer to answer the questions while "on the go". The answers to the questions are mostly simple binary yes/no questions ("Are you currently taking your medication?") and some short numbers (e.g. a*

> *four-digit security code). There is also a question that requires the entry of an item from a long list (one or more medication names). The survey is short...only about ten questions.*

This is a very specific scenario, but it is optimal for speech. One could still justify speech if a few of these situational factors were different. For example, a company might want to convey a "cutting edge" image and have little interest in the fact that touchtone could be easily used for all of its survey questions. The point is that speech IVR's might be appropriate for a range of data collection situations, perhaps more so for some than others, but the survey designer needs to carefully consider the pros and cons of the particular technology in the particular measurement context.

We have not yet touched upon the potential of speech IVRs. Touchtone technology has matured, and does not seem to be improving at this point. Although some of the technologies available to human interviewers are making CATI more efficient, the interviewers themselves are not necessarily getting any better or worse. Speech technology is relatively new, and the accuracy of the recognizers, the clarity of TTS voices, and the intricacy of the dialog designs are steadily improving. Innovations are continually changing the landscape.

For example, new architectures for dialog management are allowing speech IVR's to approach a "mixed initiative design", in which respondents can answer more than one question at a time, and move to other questions in the dialog. For example, if the IVR asks the caller "What's the departure station", the caller can respond "New York Penn Station...oh...and I have a Triple A discount". Because this technology is relatively new, development time for any given application is still considerably longer than it is for applications using traditional "system-initiated designs". However, like any other technology, cost and effort can be expected to decrease over time. For certain surveys where the order of responses is not crucial , this technology may provide benefits.

Another technology that may be helpful is called "hotword" (McGlashan et al., 2002). In certain modules, the recognizer can be set to only listen for a specific item, and ignore anything else that might be heard. The benefit here is that the system can sit quietly while the caller does other things without concern about interruption by background noise. For example, if a respondent needs to provide a social security number, cannot recall the number, and needs to get their social security card, the IVR can wait while the person gets it. It might say "Okay, I'll wait while you get that. When you're ready, just say "I'm back." Then the system will only be listening for "I'm back" and any other noise will be recognized and rejected. If a dog barks in that time, or if the person is walking with the phone and breathing heavily, the system will be less likely to make a false positive and proceed to the next question.

More subtle improvements are also happening, in the form of improved recognition algorithms. With each new release of a recognizer, the developers have figured out a way to make it *slightly* more accurate. These slight improvements are quite noticeable when one looks over decades. Clayton and Winter were working with technology in 1992 whose use today would be unthinkable.

The point is, even if one is skeptical of speech technology as useful for survey interviews, the landscape is constantly changing and it may be more viable with every passing year. It is similar to the old saying about the state of Florida: “If you don’t like the weather, wait five minutes.”

References

Bloom, J.E. (1999). Linguistic markers of respondent uncertainty during computer-administered survey interviews. Doctoral dissertation, New School University.

Brennan, S. E. (1991). Conversation with and through computers. *User Modeling and User-Adapted Interaction,* 1, 67-86.

Brennan, S. E. (1998). The grounding problem in conversation with and through computers. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication (pp. 201-225)* . Hillsdale, NJ: Lawrence Erlbaum.

Clark, H. H. (1996). *Using Language*. Cambridge, MA: Cambridge University Press.

Clayton, R.L., & Winter, D.L.S. (1992). Speech data entry: Results of a test of voice recognition for survey data collection. *Journal of Official Statistics, 8(3)*, 377-388.

Conrad, F.G., & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In *Proceedings of the 3rd International Conference on Survey and Statistical Computing* (91–101). Chesham, UK.

Conrad, F.G., & Schober, M.F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly, 64,* 1–28.

Couper, M.P. (in preparation). Technology and the survey interview/questionnaire. In M.F. Schober, & F.G. Conrad (Eds.), Envisioning the survey interview of the future. Hoboken: Wiley, Inc.

Couper, M.P., Singer, E., & Tourangeau, R. (2004). Does voice matter? An interactive voice response (IVR) experiment. *Journal of Official Statistics*, *20 (3)*, 1-20.

Currivan, D., Nyman, A.L., Turner, C.F., & Biener, L. (2004). Does telephone audio computer-assisted survey interviewing improve the accuracy of prevalence estimates of youth smoking? Evidence from the UMass Tobacco Study. *Public Opinion Quarterly, 68,* 542-564.

Deshmukh, N., Duncan, R. J., Ganapathiraju, A., & Picone, J. (1996). Benchmarking human performance for continuous speech recognition. In *Proceedings of the ICSLP-1996* (2486-2489). Philadelphia, USA.

Dillman, D.A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, & J. Berck. (2002). Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response and the internet. *Draft paper,* retrieved August 1, 2006, from http://survey.sesrc.wsu.edu/dillman/papers/Mixed%20Mode%20ppr%20_with%20Gallup_%20POQ.pdf

Fowler, F.J., & Mangione, T.W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: SAGE Publications.

Graesser, A. C. & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology, 85(1)*, 136-151.

Groves, R.M. (2005). Dimensions of surveys in the future. Paper presented at the University of Michigan Workshop *Designing the survey interview of the future*, Ann Arbor, USA.

Lippmann, R.P. (1997). Speech recognition by machines and humans. *Speech Communication, 22(1)*, 1-16.

McGlashan, S., Burnett, D., Danielsen, P., Ferrans, J., Hunt, A., Karam, G., Ladd, D., Lucas, B., Porter, B., Rehor, K., Tryphonas, S. (2002). Voice Extensible Markup Language (VoiceXML) Version 2.0. Retrieved February 19, 2007 from http://www.w3.org/TR/2002/WD-voicexml20-20020424/.

Meyer, B., Wesker, T., Brand, T., Mertins, A., & Kollmeier, B. (2006). A human-machine comparison in speech recognition based on a logatome corpus. Paper presented at the *Speech Recognition and Intrinsic Variation Workshop (SRIV2006)*, Toulouse, France. Retrieved August 1, 2006, from ISCA Archive http://www.isca-speech.org/archive/sriv2006.

Miller-Steiger, D.(2006). Interactive voice response (IVR) and sources of survey error. Paper presented at *Telephone Survey Methodology II Conference*. Miami, USA.

Mingay, D.M. (2000). Is telephone audio computer-assisted self-interviewing (T-ACASI) a method whose time has come? Proceedings of the *Survey Research Methods Section, American Statistical Association* (1062-1067).

Moore, R.K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech 2003* (2582-2584).

Nass, C., Robles, E., Bienenstock, H., Treinen, M. & Heenan, C. (2003). Voice-based disclosure systems: Effects of modality, gender of prompt, and gender of user. *International Journal of Speech Technology, 6(2)*, 113-121.

Oviatt S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech Language, 9(1)*, 19-36.

Pinker, S. (1994). *The language instinct*. New York: HarperCollins.

Pinker, S., & Bloom, P. (1990). Natural languages and natural selection. *Behavioral and Brain Sciences, 13*, 707-784.

Sacks, H., Schegloff, E. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language 50(4),* 696-735.

Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing improve survey data quality beyond the laboratory? In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (910-915). Alexandria, USA.

Schober, M.F., & Conrad, F.G. (2002). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, & J. Van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 67–94). New York: John Wiley & Sons.

Schober, M.F., Conrad, F.G., & Fricker, S.S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18,* 169-188.

Suchman, L. & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association 85(409),* 232-241.

Suessbrick, A. (2005). Coordinating conceptual misalignment in discourse and the limits of clarification. Doctoral dissertation, New School University. Dissertation Abstracts International, 66B (01), 589.

Suessbrick, A., Schober, M. F., & Conrad, F. G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association, Section on Survey Methods Research* (907-912). Alexandria, USA.

Tourangeau, R., & Smith, T. W. (1998). Collecting sensitive information with different modes of data collection. In M. P. Couper, R. P. Baker, J. Bethlehem, J. Martin, W. L. Nicholls II, & J. M. O'Reilly (Eds.), *Computer Assisted Survey Information Collection* (431-453) . New York: John Wiley & Sons.

Villarroel, M.A., Turner, C.F., Eggleston, E.E., Al-Tayyib, A.A., Rogers, S.M., Roman, A.M., Cooley, P.C.,Gordek, H. (2006). Same-Gender sex in the USA: Impact of T-ACASI on prevalence estimates. *Public Opinion Quarterly, 70(2),* 166-196.