Where does speech make sense - now, and in the future?

By
David Attwater (david @ eiginc.com)
Jonathan Bloom (jonathanb @ speechcycle.com)
Jason D. Williams (jdw @ research.att.com)

Table of Contents

Introduction

Over the past decade, speech recognition technology has worked its way into many facets of day-to-day life. Nowhere has it staked a more secure claim than in the telephony IVR space. However, certain pressures are beginning to challenge speech's role on the phone, including market saturation, the rise of customer self-service on the internet, user aversion to the technology, and a more realistic understanding of the technology's strengths and weaknesses among vendors and customers.

In light of these pressures, several questions arise. First, can we use our current, more objective understanding of speech recognition to identify where the technology still makes sense in the context of speech IVRs? Second, given that the future holds no guarantees for speech in the IVR space, in what other contexts does the technology make sense? We will attempt to address these questions in the following paper.

The Pressures Facing Speech IVRs

Touchtone IVRs have been in use since the late 1970's. The benefits to business were obvious: Automating calls cut costs in the call centers. The customers of these businesses tolerated the new automation, possibly because they had little choice; other methods of interacting with businesses (in person, via mail, fax) were limited and inefficient. The switch to limited automation also minimized the need to hold for an agent and for a given staffing level could shorten hold times for those who did require agent assistance.

Speech technology was added to the IVR interface for several understandable reasons. For example, as Ferdinand de Saussure stressed a century ago (1857—1913), spoken language is the primary means of human communication. Also, language allows for flexibility of input that touchtone does not. As a final example, if the keypad is located on the handset, entering information via touchtone and then listening to IVR prompting requires moving the headset back and forth from the head. Due to arguments like these, an attitude of "speech everywhere" took over the field of IVR design and development.

But after many years of heightened expectations and widespread growth, speech technology - and IVRs in general - are beginning to run up against realities that may threaten the future of both.

Assuming that the IVR as we know it continues into the future unvanquished, speech's role within it seems unlikely to grow and may even recede where it has been inappropriately applied. The authors have experienced customers requesting the removal of speech from IVRs, and we have even found ourselves advising against speech in some situations. From the caller to the call center managers to the speech vendor, most players are now aware of speech's shortcomings. First, it can be more than twice as expensive to install when compared to touchtone-only. Second, it usually requires more expensive, specialized people to support it like speech scientists and VUI designers. But these issues would be acceptable if the incremental benefit over touch-tone was large enough. And that is where the largest problem comes into play: In many cases, speech technology can actually hinder the caller experience instead of improving it. Because of the uncertainty inherent in speech, keeping a stable user interface in response to things such as no-matches can lengthen individual interaction. This is in contrast to the predictable and somewhat binary nature of touchtone. As a consequence speech can shorten the average handling time (AHT) of successful calls, but be slightly less effective overall. Take as an example open-ended "How may I help you" style questions. Human interlocutors have to assume basic common ground while engaging in a conversation (Clark & Brennan, 1991; Clark, 1996). For example, they must assume that they share a common language, a common set of beliefs, etc. When an automated system suddenly asks “How may I help you” very little common ground can be assumed. Because of this, callers often err on the side of caution and assume the system shares nothing with them, especially not deep semantic and pragmatic processing. The caller utters short utterances with little content and so the system is required to fall back to the menu structure it was originally designed to avoid. In light of these challenges, the simplicity of touchtone menus begins to look more attractive.

But can we really make the assumption above? Will IVRs –touchtone or otherwise - continue to play a role in customer interactions? Perhaps not. Arguably the starkest reality facing the IVR is the continuing rise of the internet. Customers are turning more and more to company websites in order to complete simple queries and transactions. For example, responsibility for completing a train reservation used to fall to IVRs and agents. Now, a customer can complete a train reservation via the web. This is preferable for customers because it can be completed at one’s own pace, movement back and forth in the interaction is more obvious, and information presentation is static and parallel (as opposed to ephemeral and serial). This is also preferable for businesses because it costs less for them to support web interactions than to support calls into call centers. Because of this major shift in the customer-business relationship, internet aware customers are not reaching for the phone until they know that they have a situation too complex for automation and that they need to speak to a human being. As internet usage increases for a given industry sector this leaves little room for the IVR as we know it.

Where Speech Makes Sense: IVRs

In the near future speech IVRs will certainly continue to exist. We believe that speech technology does provide benefits in IVRs, but that vendors and customers need to be more subtle and conservative about its usage. As the author Jared Diamond wrote, we are caught up in a culture where “invention is the mother of necessity.” It is our responsibility as advocates for the caller to steer all parties away from that view of speech technology. Just because we can do something with speech does not mean we must do that thing with speech.

So where does speech make sense? There is one specific context in call flows where speech seems preferable to touchtone, and that is in the case of “item lists.” By item lists, we are referring to a list of items too long to be presented to the caller as a menu. Some common examples of item lists include city names in flight reservation applications, employee names in auto attendants, and company names in stock quote applications. The touchtone alternative to speech in item lists is cumbersome, often requiring callers to select letters off of the keypad that correspond to the first several letters of an item.

In addition to item lists, speech can make sense in the case of “How may I help you” questions but only if the caller’s expectations can be set properly. As mentioned earlier, these contexts can trip up callers because they are not sure what common ground they share with your system. If you provide an example of what can be said in your prompting, callers often fare better in crafting an utterance and proceeding successfully to the next step.[1] Specific scripting of the question is also very important. For example, when creating a call router for Microsoft in 2001, one of the authors saw callers craft utterances of a more appropriate length for recognition performance when the word “briefly” was used in the prompting (e.g. “Briefly describe the reason you’re calling, for example ‘I want to order new service.’”) Between the example and the use of the word “briefly,” short, ambiguous utterances and long, indecipherable utterances are minimized.

[1] The risk in examples is that they tend to lead to callers parroting the example even though it does not reflect their call reason.

One can also argue for speech in cases where the calling population is expected to be calling from a cell phone. Many phones are designed with a force function, locking the keypad during a call or even hiding it. In such cases, speech may be a preferable modality. However, cell phones are often used in noisy environments, so the benefits of speech must be weighed against the risk of elevated no-matches.

Where Speech Makes Sense: Elsewhere

Regardless of what the future holds for IVRs, the speech industry need not worry. Speech technology has many other reasonable applications. In our discussions, we found three key motivators for employing speech as a modality, either alone or paired with others. Note that these motivators are not mutually exclusive:

Small form factor

Fitts' Law (1954) implicitly suggests that the effort required for a person to touch (e.g.) a button on a smart phone screen is a function of the size of that button. The smaller that button is, the harder the person has to work to tap it. We now live in an age where computer screens are becoming smaller and smaller on average. The shrinking form factor of computing devices certainly offers many conveniences in terms of mobility and productivity. But few of us would even consider writing a book using a Blackberry because of the incalculable effort the interaction would require of us. Many of us will fill out forms on our handheld devices using a keypad and we will often write short emails. However, speech can provide a faster method that demands fewer fine motor movements. Even a simple web search can be facilitated using one's voice. Speech is a reasonable supplement as screen sizes shrink and touch becomes too burdensome for some tasks. Of course, issues of social etiquette (dictating in public) and ambient noise still present challenges, but not ones that are insurmountable.

Eyes-busy/hands-busy environments

People often find themselves in situations where use of their eyes or hands is restricted. The most obvious example of this is driving a car. Strong evidence exists that, along with other distractions, speaking on a cell phone while driving increases the likelihood of an accident (National Highway Traffic Administration, 2010). Meanwhile, results on the safety of placing phone calls using a hands-free device are less clear. Assuming for now that hands-free and eyes-free solutions are safer than dialing and holding a phone to one's ear, speech recognition and text-to-speech become viable means of interaction. Voice dialing has been in existence for some time now and seems likely to stay with us for some time.

Other eyes-busy/hands-busy environments do exist that lend themselves to speech. Video games offer another example. In addition, one author is aware of a software company that has developed speech macros to assist their AutoCAD developers in the completion of more repetitive tasks. These macros allow the developers to continue working with mouse and keyboard uninterrupted.

Accessibility

One of most obvious uses of speech technology is the case where there are few alternative methods of interaction. Individuals who have limited use of their hands – for example in cases of repetitive strain injury (RSI) – can use speech for both dictation and command & control (e.g. moving the mouse around the screen, formatting documents, opening and closing windows, etc.) on any sized form factor. The most dedicated users of Dragon's NaturallySpeaking® software are those with RSI and similar impairments. They have created very active online user groups that share complex macros for completing everyday tasks. Some of these individuals have gone so far as to incorporate foot pedal input in addition to speech macros.

These three motivators – small form factor, eyes-busy/hands-busy environments, and accessibility issues - are likely to drive emerging applications of speech technology outside the call center. To be sure, there are some other applications of speech that do not neatly fall under one of these. For example, language learning software like Rosetta Stone® can benefit from users speaking a new word or phrase in the target language and having a recognizer grade the user on pronunciation. In an example of a truly multimodal interaction, in which two modalities are not only used but used simultaneously, speech can be used to add meaning to gesture. An early example of this was Richard Bolt's "Put That There" project in which users could get assistance from a computer placing objects on a map. When asked by the computer where to place an object, the user could say "there" while pointing. The gesture disambiguated the deictic utterance.

Conclusion

Speech still provided benefit in IVRs, especially in the case of item lists. Speech also makes sense for open-ended questions, but the designer needs to ensure callers have enough context to provide meaningful responses. If the wording of those prompts is mishandled, then speech can be a detriment when compared to touchtone.

We also identified several places where speech makes sense outside of the IVR space:

(1) small form factors such as smart phones

(2) eyes-busy/hands-busy environments such as phone dialing while driving, and

(3) accessibility needs such as command and control of one's computer desktop.

Speech also serves an important role in specialized applications such as language learning software. Finally, an interesting future direction is to augment touch-based interactions with simultaneous speech input.