I Should Make My IVR Sound Natural, Right?

By Jenni McKienzie, David Attwater, Jon Bloom, Phillip Hunter, Greg Simsar, Louise Tranter

Introduction

As VUI designers love to say, it depends. "It depends" is a perfectly healthy attitude to have in design in that there are very few things that work across the board. But let's define what it depends on. As to naturalness in an IVR, it depends on what you mean by naturalness. Let's look at various levels of the definition of naturalness and come to an agreement of just what we're talking about and then address what we should aim for.

At one end of the spectrum is the definition where naturalness means the IVR is able to understand and respond with completely unconstrained speech. While we all agree this would be nice, the technology just isn't there yet. Next to that is where the IVR stays completely truthful to emulating human-to-human conversations. In the early days of speech recognition in IVRs, many a designer strove for this ideal, albeit sometimes in an overly-stylized fashion. Some still may. The opposite end of the spectrum, which we're hopefully far away from now, harks back to the early days of DTMF IVR with phrases like, "invalid response." Where the majority of the industry has landed is that the first priority is the success of the caller. At times this means varying from a strict human-to-human model. After all, we are talking about a human-to-machine interaction. But where naturalness still comes into play is that beyond making the system work for its users, we as designers must build on, take advantage of, and adhere to the innate rules of human conversation.

Our Definition

A short digression into what naturalness is not. It is not a conspicuous or over-the-top persona that attempts to emulate a character that the designer wants the caller to remember. While this may work for a particular situation, it is definitely not necessary to achieve a natural interface, and in fact can sometimes create just the opposite. Lest we be accused of throwing the baby out with the bath water, well-defined personas can be more low-key and transparent. In this more subtle use of persona, naturalness is not determined by style. For instance, an interface may be called polite, with pleases and thank yous scattered liberally around, yet the two concepts are independent. It is possible to design a system that is exceedingly polite yet violates the basic constructs of conversation. Conversely, a system can conform beautifully to these constructs without ever apologizing or saying please.

A quick example to illustrate our definition. Consider how you speak to your dog. Your speech is not the same as it would be to another person, it is shaped for the dog and what he can understand. And yet it is modeled on and borrows from human-to-human conversation. Every human-to-blank conversation does. The basic constructs of language remain the same. And yet we're all very good at adapting to what is in that blank: a dog, a small child, a machine. At the heart of this adaptation is the very human ability to anticipate and adjust to the capabilities and beliefs of the other conversant. In all conversations we are continually forming a theory-of-mind of the beliefs, capabilities, hopes, and desires of the person or thing that we are conversing with. Machines have virtually no capability in this regard, but a person engaged in a human-to-machine conversation certainly does, and so should designers!

What We Know About Conversation

So what are these basic constructs that designers need to be cognizant of? Let's start with Paul Grice's maxims of conversation: quality, quantity, relation, and manner.[1] Quality means be truthful. Quantity means say everything you need to and nothing more. Relation means make it relevant. Manner comprises several elements: clarity, brevity, and order. All of these are excellent guidelines for writing prompts. What we must never forget is that we are writing spoken language, and that spoken language leads us to a couple of other principles. The first of which is turn taking. If we desire to communicate successfully and empathetically, we take cues from our conversational partner as to when it's our turn to speak. It is imperative that designers are aware of this when crafting prompts. Another vital construct is that of discourse markers, the little transition words that act as acknowledgments, indicators that a shift is about to occur, sequencing markers, and countless other functions. Strip these out of a conversation and it becomes very difficult to follow and feels stilted and emotionless. So by naturalness, we mean take advantage of all these constructs and being aware of them.

IVRs have two basic functions: relay information and elicit responses. The degree to which you can model human-to-human conversation varies between the two. The former is where we can really leverage human-to-human constructs, using, for example, tapering, colloquialisms, and following the Gricean maxims closely. An example from travel. The caller indicates they want to book a new hotel reservation and hear the following.

> *Just so you know, all available ABC rates are published on the website. Agents will be more than happy to help you make the reservation, but have the same rates as the web.*

Here you see a discourse marker in "just so you know" that prepares the caller for a little extra information. There's a little tapering in going from "website" to "web." It's truthful, giving the caller a

[1] Grice, P. (1975). Logic and Conversation. In Syntax and Semantics, Vol. 3, Speech Acts. Eds. P. Cole & J.L. Morgan. New York:

level set. It's relative, addressing a common reason for calling. All in all, it sounds like something a person would say.

A second example.

> *Okay. You have three transactions on your account in the last two days. You used your credit card yesterday at Borders for twenty seven dollars and forty cents... today at Piggly Wiggly for twelve sixty five.. and again at a Marriot hotel for three hundred forty two.*

Notice the leading discourse marker "Okay," the relevance of only tracing back a couple of days, and the tapering of the transactions.

Getting the Recording Right

This is an excellent time to discuss voice talent delivery. In the above example, a little emphasis will have to be placed on "yesterday" and "today" to help the caller conceptualize the list structure. The best written prompts can be abject failures if the voice talent doesn't deliver them right. It is the designer's job as the voice talent coach to make sure all prompts are delivered as intended. The right amount of emphasis has to be on the right words and the right parts of words. Consider this example in Spanish that corresponds to an English prompt asking if the reservation is international or domestic.

> *Para reservaciones nacionales, oprima el uno. .... reservaciones internacionales … dos.*

There was simply no getting away from the shared word-form in "nacionales/internacionales". This, combined with the tapering, demands that we get the intonation just right.   The voice talent had to over-emphasize the "inter" part of the latter to emphasize the contrast and make sure that callers hear the difference.

Pauses have to be appropriate too. What constitutes appropriate? Again, the underlying rules of turn-taking and the limitations of human memory remain the same for human-to-machine dialog, but the demands of the user interface may dictate less "natural" use of these rules. Take for example a listed menu structure. For longer or more verbose lists there needs to be enough of a pause for the caller to process each item, then either respond to it and barge in, reject it as definitely not what they want, or remember it as a possibility. To do this effectively, pauses need to be somewhere between about 500 and 1000 milliseconds depending on the complexity of task or language used. The use of such pausing is "natural" in situations where people normally choose to communicate with lists (consider the waiter reading the specials in a restaurant). What is not "natural" is our need to frequently rely on such lists in order to create stable user interfaces. Such lists are more likely to be replaced by spontaneous conversation in a human-to-human conversation with pauses closer to 150 milliseconds. Thus the need to use lists demands longer pausing than spontaneous conversation normally requires.

For other types of prompts, a pause might need to be minimized to discourage callers from speaking in it. And pauses have to go in the right spot. A pause in a wrong spot can be misleading to callers.

Enunciation matters too. Many voice talents are used to over-enunciating, making sure to be crisp and get every consonant. The word “eight” may sound more like “eigh-tuh,” which is definitely not natural sounding. It won’t work at all to use casual speech and colloquialisms if they are over-enunciated.

How to Elicit Information

Now look at eliciting responses. This is where designing gets tricky because we have to rely on the technology, both the recognition engine and the grammars we’ve written, to process the response, as well as human expectations of conversation and machine capability and behavior. The prompt has to be written to strongly influence that response and ensure that we can as accurately as possible predict what it could be. Two examples here.

> *Say or enter your 10-digit home phone number.*

If we were modeling a human-to-human conversation, this prompt would be as follows:

> *What’s your home phone number?” or even “Can I get your home number?*

And several years ago many of us probably would have written one of those, and some of us definitely did. But there are a couple of things playing into this that make the more human-seeming strategy faulty. The first is that callers have a modality choice. They can in fact either speak their response or enter it via DTMF. Since humans don’t decode DTMF, they don’t prompt for it in their speech. Our IVRs can offer it, so we as designers can leverage it. Many callers are more comfortable with DTMF for digit strings for reasons of either accuracy or security, especially for sensitive identifying information. The second factor here is possibly setting the theory of mind of the caller unrealistically, attributing more intelligence and human-ness to the IVR than exists. This can lead the caller to start providing more of a response than the machine is capable of handling. It’s entirely possible that the error handling would function as desired and the conversation would recover, but why take two or more turns to do what can easily be done in one with the right prompt? This is a perfect example of where abandoning a pure human-to-human model can result in a more effective interaction.

A second example from a technical troubleshooting application that is trying to elicit a yes/no response. Contrast the two approaches.

> *Do you see a green light on the front of your modem, just next to the power button?*

> *On the front of your modem there should be a green light next to the power button. Do you see it?*

It’s probably safe to say that almost anybody would say that the first sounds more like human-to-human conversation. But any designer worth their salt will tell you the second one will succeed in eliciting the “right” response more often. Some might interpret this as sacrificing some of the Gricean maxims (namely order) as they apply to a traditional conversation between a customer and a call center agent. At the very least the first is more colloquial or conversational. But this is done to ensure that the caller doesn’t short-circuit the question. If two people are talking and that were to occur, the human could very easily identify that the respondent didn’t hear the whole question and verify that the appropriate answer was indeed given. Without carefully modeling caller knowledge during turn-taking this is much harder for the speech IVR to do, so it is better to simply make sure it doesn’t happen in the first place.

Is it Naturalness?

The naturalness question has taken on other names and guises over the years. For example, designers often debate the prudence of using first person in speech IVRs. The first person debate is really just a small example of the larger naturalness question, and academic arguments exist for both sides. There are designers that feel that the use of first person leads to overly high expectations on the part of the caller, that modeling human-to-human conversation can lead to overly verbose responses that the grammars can’t handle. On the flip side, because we’ve already established that our human-to-machine conversation builds upon the constructs of human-to-human conversation, it is perfectly natural and acceptable for the former to use first person since the latter obviously does. In many cases trying to avoid the use of it leads to prompts that are more awkward and unnatural. Just like we dislike the “royal we” in our human-to-human exchanges, so attempts to refer to “the system” sound odd and distract the caller. One could also argue that a human is so heavily implied by the use of a human voice, that avoiding first person in prompt wording is a strategy that is too little too late. But the bottom line is that both can work. The two important things are to be consistent and to avoid the first person plural unless truly referring to the company as a whole. We’ve all come across people that use “we” instead of “I,” and I think we’d all agree that they annoy us.

Let’s assume now that naturalness is about conversational constructs more than strict emulation of human-to-human speech. What are some examples of where abandoning that strictness leads to more effective prompts? Most designers would agree that the double structure of, “For A say A” is not only unnatural but ineffective, ugly, and not using the full potential of the medium. But what about a variation of it to lighten cognitive load? There are times when the choices to be presented to the caller need a little introduction or explanation. You want to prep them before eliciting the response. For this situation, a structure of, “I can do A, B, or C, so say a, b, or c” can be very effective. Here’s an example.

> *You have three choices for the days your newspaper is stopped. I can credit the days back to your account. You can donate them to our newspapers in schools program. Or your carrier can deliver all the missed issues when you return. So, say credit my account, donate them, or deliver them.*

The brief phrases at the end, the “a, b, or c” part, aren’t clear enough by themselves. But if you use the phrases in the front part, the explanation or the “A, B, or C” part, they are clearly too long and produce cognitive load issues. The balance between clarity and conciseness is to be clear upfront in the description, then concise in the utterances that are modeled.

Statistical Language Models and Naturalness

Statistical language models (SLMs) afford us their own challenges with naturalness. Here's a situation where we expect the caller to respond more naturally and from a broad array of more specific information than in a directed dialog situation. There are a couple of issues here. The first is that we've managed to train people to respond to directed dialog, via DTMF or speech, fairly well. Varying from this and not giving them the guidance of a menu leads to an increase in uncertainty and thus cognitive load. They're not sure what to say because they aren't sure what the system expects. The second is closely related. If this is the first question they encounter in an application, they are possibly even more thrown off by the type of question because they haven't yet engaged in the conversation. So how do we make it easier for them? Address the second point first. Put something before the SLM, preferably an easy to answer yes/no question or a simple menu. That gets them engaged and playing along but doesn't wall off the possibility of expanding the dialogue type.

Going back to the first problem, much trial and error and experimentation has shown that you have to give callers examples. This is another example where deviation from natural human-human conversation is used to manage caller expectation and theory-of-mind. But how many examples? How long should they be? Do you go with high, medium, or low hitters? Many more experiments have been done on these issues. These are interesting in the realm of naturalness, because in a human-to-human conversation, you simply ask the "How may I help you" type of question and move along. But that just doesn't fly as well as we would like in human-to-machine interactions, thus the research into what's the more effective. Various studies have been done on various applications with varying results. (Remember our "it depends" mantra?) One study showed that the best results came from using two examples, medium hitters, with short verbiage. So give an example of "lost internet connection" rather than "I lost my internet connection." But another company has found for their application two examples is disastrous, that one is the more effective. There is much still to understand here but, the answer will almost certainly lie in understanding caller's beliefs about the capability of the machine and the role that examples play in priming in modifying these beliefs.

Menus and List Selection

What's the most natural way to structure choosing things from a list? Is there such a thing? So far we've looked at SLMs (where we've decided we use a most unnatural construct of two examples) and a double prompted situation. A lot of designers argue that starting with a single yes/no question is the best way to engage and get started. But not every situation lends itself to that. Suppose you have six functions that you want to include in your list or menu. What's the most natural way to structure it? A single menu with six options (which can be made to work just fine, but that's another topic)? Maybe a sub-divided menu of three options plus "help me with something else" followed by the second three? What about the highest volume single option as a yes/no, then the following five? The right answer has nothing to do with what's the most natural. It has to do with what's the most effective. And the answer to that is data-driven. If one of those six options covers more than 50% of callers, then start with a yes/no just about it. If the first three options cover 85%, try the two-level menu. If they're all pretty much equal, then present them all at once. Or even better, try all three approaches and see what works best. Test them all, but never forget basic human psychology, the passage of time, and the fragility of human memory.

Conclusion

Returning to the question posed in the title, IVRs should imitate traditional human-to-human conversational constructs unless it is more effective to do something different. Even then, basic conversational interaction principles should almost always be adhered to unless the context specifically supports straying from it. At the end of the day, human-to-machine conversations are not identical to human-to-human and should differ, but not by much. The innate rules of conversation still apply.