

SOUND IN THE MACHINE

Written, produced and edited by Brian Snead



EPISODE TITLE

Machines that Talk to Us

TIME

15:07

DATE OF WRITING

10 November 2008 – 26 November 2008

PUBLICATION INFO

Date: 28 November 2008

Number: 3

Location: Sound in the Machine.org:

<http://soundinthemachine.org/2008/11/28/machines-that-talk-to-us-2.aspx>

SUMMARY

From talking self-checkout machines, microwaves, child learning aides, screen readers, street crossing signals, and automated museum tour guides, we are surrounded by machines that talk to us. But they are a long way from holding a conversation. The best anyone has been able to come up with so far are domains to which the human speaker is extremely limited. But the technology that makes the current machines talk is varied and interesting and will underpin those machines decades into the future with which we will be able to have meaningful interaction.

KEYWORDS

Text-to-speech synthesis, self-checkout, screen reader, auditory display, sonification, narration, sound at the user interface

(Some) SOUND SOURCES

[Vox]: "[There's a Place Called Tomorrow](#)." YouTube. [obocat](#) Channel.

[C3PO]: "[Robosapien - C3PO banter](#)." YouTube. [tankd0g](#) Channel.

[Russian]: [NHK World Russian Podcast](#).

[Atari]: The Free Sound Project. [Irrlicht](#) User.

--"[press play on tape](#)"

--"[game over](#)"

--"[extreme terror](#)"

--"[rock and roll](#)"

[UK Telephone message]: "[1471 a.wav](#)." The Free Sound Project. [Benboncan](#) User

[Talking Microwave]: "[Hamilton Beach Talking Microwave - 87108](#)." YouTube. [Venkateshnt](#) Channel.

SOUND IN THE MACHINE

Machines that Talk to Us



published 28 November 2008 at soundinthemachine.org.

By Brian Snead

TRANSCRIPT

You're listening to [sample "Sound in the Machine." (LH Michelle on Adobe 8)]

In the 2002 film *The Time Machine*, there is a fictional demonstration of perhaps the pinnacle of talking machines. [sample]. Vox represents the apotheosis of human intellect, later in the film becoming the mysterious "ghost" conscience of humanity after it has managed to rip itself apart.

Let's contrast this with a non-fictional situation in which we may find a machine talking to us. [sample "Welcome Valued Customer!"] Oh yes, the self-checkout lane: [sample] It talks to us, we don't get to talk back, not in public anyway though I'm sure you've felt, as I have, "inspired" to respond when hearing for the 3rd time: [sample "place item in bag"] after its been in the bag for at least 20 seconds.

Why would we program a machine to say stuff like [sample "123 rewards master card"] Or [sample "thank you for shopping with us"] in the first place?

Allegedly, the voice-capable automated checkout is a great tool to get shoppers through the payment process more accurately. And they also are a perfect opportunity to reinforce a "store experience." Unlike a human employee, the machine never tires of repeating the same slogans in a super-duper-peppy voice. Can't you just see the radiant face of the cashier, her gleaming bleach-white teeth and vacuous expression as [sample "Welcome Valued Customer!"] trips off the tip of her tongue? I'm sure anyone who has ever worked in a customer service job and remembers the hours of videos you watch your first week in the break room can certainly see and hear that corporate-approved radiance. To be sure, voice automation is perfect for old-school advertising wanks.

But apart from brand reinforcement, automated voice machines can assist the visually impaired and those with cognitive disabilities that make understanding visual text and computer metaphors difficult. Voice automation is also a great way for an application to remind you to do something important, such as my usual ATM, which reminds me to take the receipt that I forget every time. It can also encourage multi-tasking. If you have the option of doing other things while information is spoken to you, such as scratching your feet while the answering machine plays your messages or listening to the news on the way to basket weaving class or being alerted to when [sample "You've got mail"] while reading a book, you are more efficient. Sometimes having to look something up takes too much time or is impossible. Like, as teenagers, we called in to Moviefone. There was a voice on the other end that told you all the movies playing in your area. Now Moviefone is online and Google offers a phone service called Goog-411. Go figure.

Answering machines, MP3 players, and our inboxes are generally helpful machines that talk to us. But there is a dark side to voice automation, darker than any check-out lane: [sample] It's so common to get the dreaded voice menu that more and more advertisements have been exclaiming, "talk to a real live person who can assist you with your call!" Its not the recorded voice so much that annoys us, it's the fact that its basically idiotic. It talks to you but you can't talk back. It's a human voice to our ears and we expect it to understand us and to be marginally intelligent. It never is, which can be bad news for the real person who gets us after we've tried every combination of numbers to get a "real live person." And then we have a euphoric moment like this: [sample "C3P0"]

Voice automation technology is not yet to the point where machines can understand any spoken input, compute a response and deliver a natural response. But it's still all around us. Take a likely scenario, say, my journey to the office each day.

I leave my house and soon come to an intersection. I wait for the light to change, then: [sample] which is telling me which road I can cross. If I were visually impaired, this may be my only way of knowing when and where. I would be further guided by the alternating beeps on both sides of the intersection. I get to the train station. When my train arrives I board to a voice telling me: [sample 'this train bound for'] This is crucial to know if you don't manage to see the front of the first car as the train arrives.

As the train approaches each stop, I hear [sample 'next stop'], excepting that it gets slightly annoying to hear this at every single stop two times each day, this is a great way to help visitors figure out where they are going. It is loud and to-the-point. But the conductors must hate it because they usually turn it off halfway through. Then they read the list of attractions and say a lot of unnecessary things about collecting "personal" belongings and thanking me for riding the train. Usually the roar of the train is so great that it all sounds like this: [sample conductor covered by train] The microphone is crappy, there is noise in their compartment and they are often distracted by other things, ya know, like driving the train. It would be better if they would just let the recording play: its loud, its too the point, and it doesn't have a speech impediment.

So, back to my journey in to the office. After exiting the train, I make my way over to my office building and to the elevator. I step on the elevator, hit the button and hear [sample "going up"]. Heading down later for a coffee, I hear. [sample "going down"]. The interesting thing here is the intonation or 'pitch' of the voice. When you are going up, the voice goes 'up.' [sample "going up"] When you're going down, the voice goes 'down.' [sample "going down"] The rising pitch is particularly strange because its the pitch we typically use when asking a question. As the heavy doors slowly close, trapping me inside, [sample- doors closing], I want the elevator to sound like it knows where its going.

This brings me back to the human-voice-in-a-machine thing. I hear human voice, I think intelligence, discourse, interaction. But just about every example I've played on this podcast was just some machine playing recordings after somebody pushed a button. I don't find that particularly spectacular. But, using sound to communicate information or for reinforcement is helpful. And using natural language instead of beeps and whistles has lots of advantages. Thing is, voice synthesis technology is not far enough along to be useful in most situations.

Broadly, there are two ways of making machines talk to us. We can record pieces of a message to be strung together and played on the spot by the machine or we can program the machine to figure out how best to pronounce a given message.

Here's the answering service again as an example of pieces of recorded messages strung together. Listen to the slight difference in sound when the voice says the numbers: [sample]

All of those times were recorded separately and were then automatically selected by the software based on whatever time of day it was. This system uses very long segments with a few smaller ones thrown in. But we can record people making the smallest possible sounds of language and put them by the hundreds or thousands in the machine's database. There are all kinds of variations in between the extremes recording every possible sound and recording a set of a few hundred sentences that play on command. This approach, which is called concatenation, is really interesting, especially in those cases in which the machine's database is filled with very small parts of speech. In the most sophisticated of concatenation systems, the machine has to turn all of these tiny pieces into something smooth and humanlike.

To get a grip on this challenge of making something intelligible and smooth out of tons of tiny pieces, think of language on the level of sound, not on the level of writing. In most utterances, there are no breaks between words. It's just one long stream of sound. Listening to English won't prove this to you. To hear it in action, check out a language you may never have heard before. [sample Russian] Since I don't speak Russian, I have no idea where individual words are because it sounds like a never-ending stream of (very cool) gobbelty-gook. Now, Japanese: [sample] See, this is what the machine's gotta do: make one long stream out of a bunch of pieces.

I've kind of described the ideal situation, at least to me, where the output is really synthesized from minuscule fragments. But when we hear concatenation systems, we mostly hear stuff that is just long phrases recorded into a database by a human. The results from this are usually good, obviously, but are rather canned-sounding. The machine is limited to the size of its database and how fast its processor can assemble all the pieces. More of either is more time-consuming and expensive to develop. So the majority of what we hear in commercial applications of concatenation systems is basically as good as it's gonna get for now.

I'm rather more interested in the machines that make the language out of the rules of pronunciation. A good example is the screen reader. A screen reader is a program that converts text on a computer screen to speech. Here to read the next few lines, Microsoft Sam on the Adobe screen reader:

[sample “This is Sam’s voice, not a manipulation of mine. Sam’s software is figuring out how to vocalize what I have typed. All Sam knows are rules. No words. Sam saying this to you is a bit like holding a conversation with an educated citizen of ancient Rome using only a Latin grammar book.”] That’s good stuff. So I like Sam, as well as his colleagues Michael and Michelle [sample Michael: “hell yeah, that’s my boy.” Michelle: “and damn if you ain’t fly, Snead.”]

These voices have come a long way from what I remember computer voices sounding like when I was a kid. Here’s an example of an Atari ST home computer around the mid-80s: [sample “press play on tape”] Hard to understand, to say the least. How about: [sample “game over”]?

Sometime in ‘86 or ‘87 I got a Speak and Read and Speak and Spell. These were handheld [sample- turning on sound] learning aides that you gave to nerdy little kids. And they were totally rad. [sample] The days of the Speak and Read and Speak and Spell are probably behind us now, and have been replaced by games like the LeapFrog Tag Reading System. Unlike the Texas Instruments stuff, which used a rules-based approach, LeapFrog uses something like the concatenation approach, where entire segments play when the child activates them. [sample]

I’m not so sure how “educational” that is, but maybe that’s just the crusty old man in me. There are machines talking to us everywhere. Usually they’re novelties, such as trash cans that burp and say [sample “burp...thank you”] and narrations on websites [sample] Some have genuinely practical applications, such as [sample] the talking microwave, who’s also, [sample “ready to cook”], by the way. There are also learning aides, screen readers, street crossing signals, and automated museum tour guides. And I would be remiss if I didn’t mention the GPS sitting on your dashboard giving you hard and fast directions with an attitude. [sample “du blöder Mann”]

For the time being, what’s most lacking are machines’ ability to comprehend what we say to them. The best anyone has been able to come up with so far are domains to which the human speaker is extremely limited. Talking machines are a long way from holding a

conversation. And they are a long way from making us think they feel anything, no matter how perky those self-checkout lane voices are.

But we can still dig that dream and hear this technology get better and better in the coming decades.