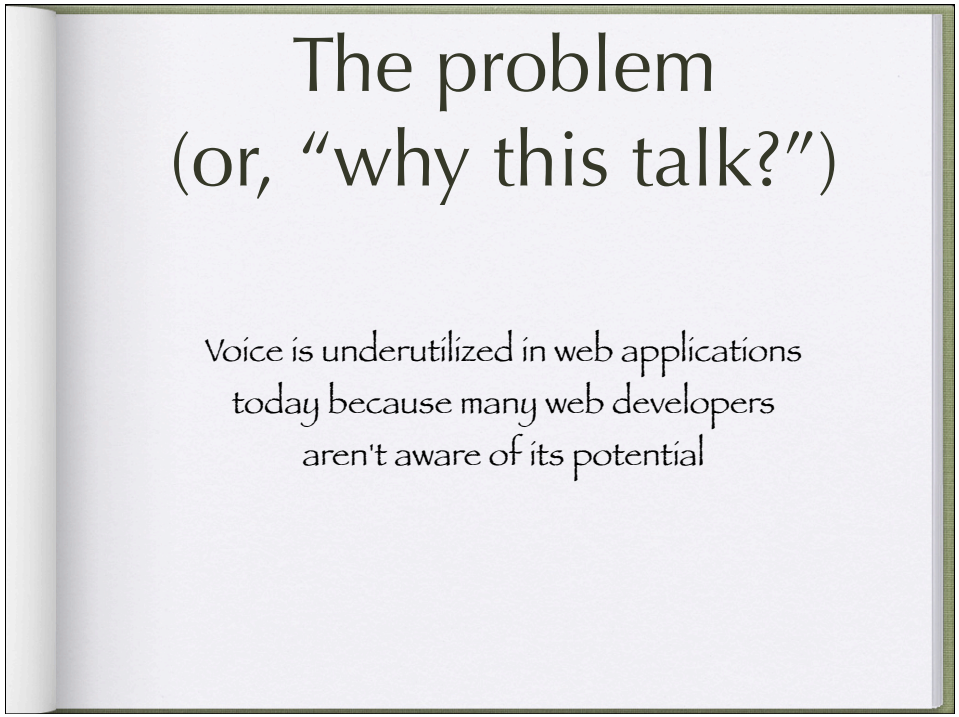
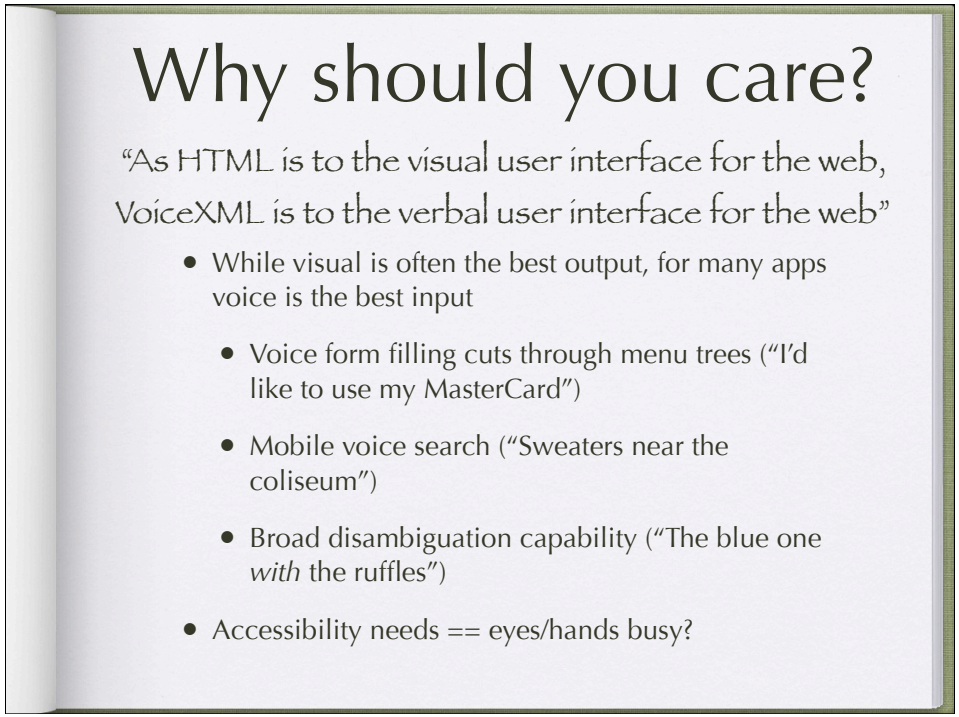
Voice Browser Working Group

Its Goals and You

12 Feb 2010



I am here because we have a strong belief that voice technology is not utilized broadly in Web applications today because many Web developers aren't aware that it exists or what can be done with it. I'll talk about why that is in a moment, but this is something we want to see change. W3C is the expert group for Web applications in general, so we believe this is the right group to target.



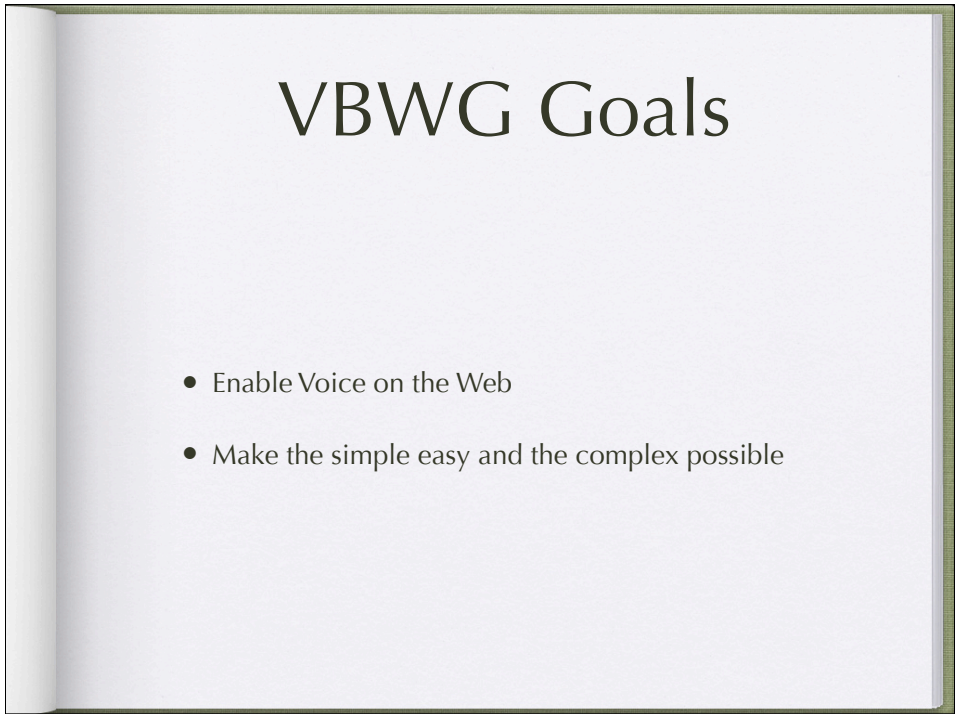
Everyone knows what HTML is for -- it provides a visual UI for the Web. VoiceXML is similar, but for aural user interfaces for the Web.

We all know that a visual output is quite rich, quite useful, but there are many applications for which mouse/touch input isn't the best, and in fact voice may be better.

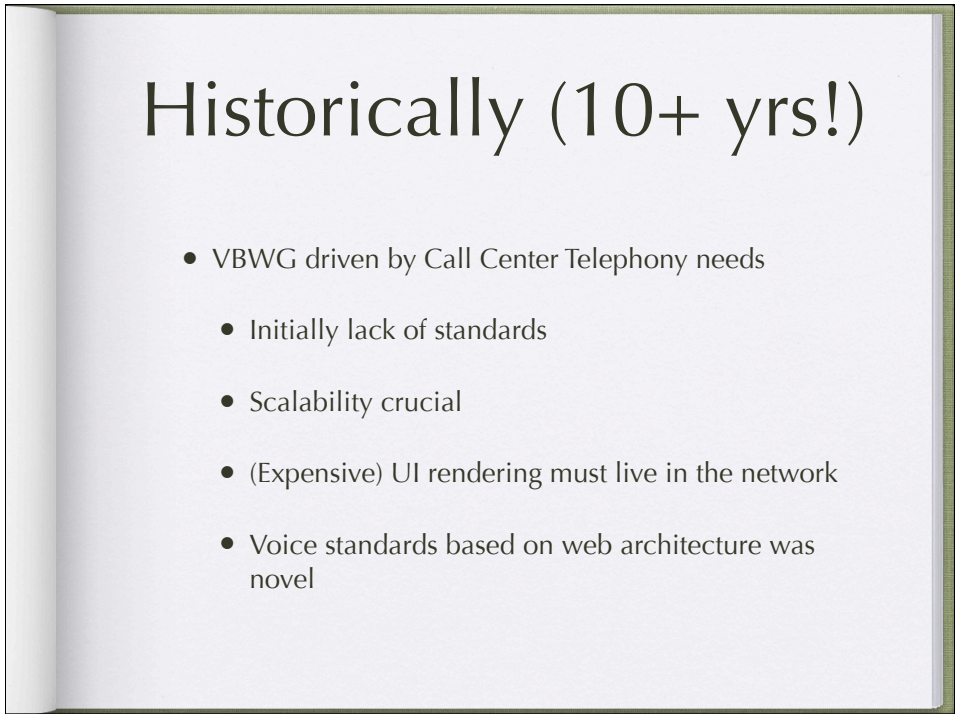
Voice does an extremely good job at cutting through form filling. For example: let's say you're purchasing something online, you find what you are interested in, and you are presented with a billing/payment screen. With a visual UI, the options are presented in a fixed order, often across multiple pages in order to avoid visual clutter. Often if you wanted to use a particular form of payment, you can only know that after wading through multiple pages. It's possible with a voice interface to allow someone to present that upfront. Even though you're being asked what item they're interested in, you can still say "I want to use my Discover card" -- it's a great advantage to the customer to do it right up front.

Mobile search is another area. Localized search on mobile devices, one of the challenges is that though people text frequently they're not conducive to typing lots of information. For instance, say you want to find a sweater, not at a particular store, just where there are sweaters, and because of where you're heading you want just sweaters near the Coliseum. Voice is very convenient for this kind of search refinement.

The last example is similar, voice provides a lot of opportunity for disambiguation. e.g. browsing on a site, where you have an option to look for something by color or by style, and really what you want is both. It makes sense to provide simple categories for people doing a visual search, but it is convenient to be able to refine and disambiguate. You may be at a screen that shows you the blue search, or another showing the search with ruffles, but blue with ruffles is what you want. One interesting thing we have discovered is that often the needs of the accessibility are very similar to some of the needs of someone whose eyes and hands are occupied. This is something voice can very much help with, on both the input and output sides.



In the Voice Browser WG we have 2 main goals:
We want to make sure that voice technologies are used on the Web
The second goal is 'Make the simple easy and the complex possible'. We use this phrase all the time in the VBWG. There is a perception that voice technologies are complicated. It's difficult to walk that line between power and complexity. This is relevant to some of the initiatives we've had.

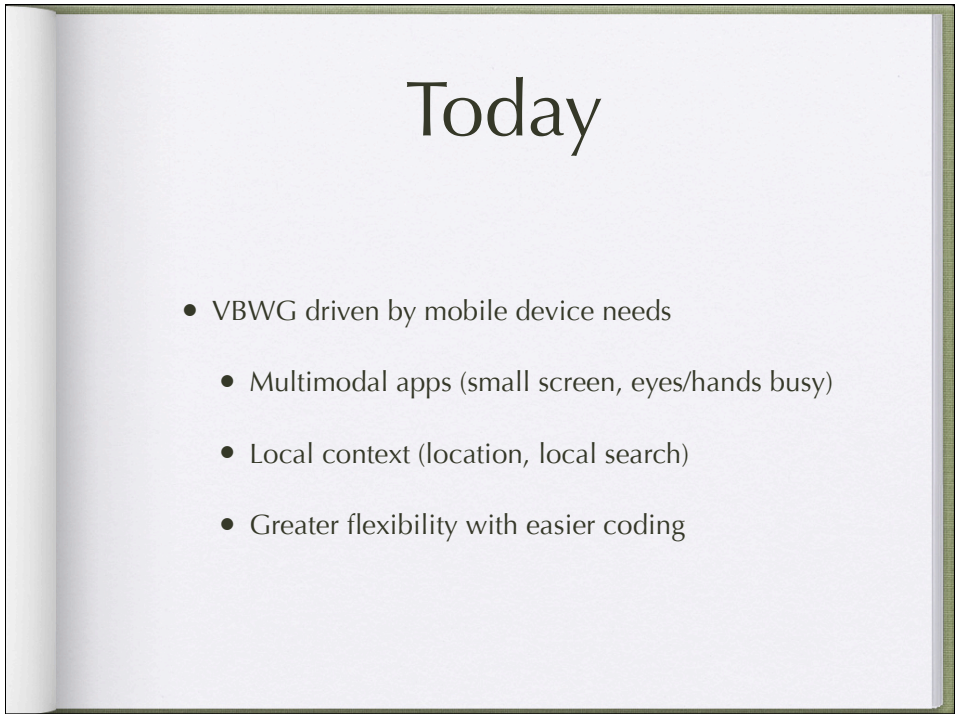


Web developers may not be aware of what our voice technologies can do. We're one of the older groups at W3C. Our group has tended to shrink and grow over time, but it's often quite large. Originally the needs of the WG were driven by the needs of call centers. Call centers are the places you call into when you want support or make account queries, etc. These operations are usually large and expensive. There was a notable absence of standards. The Java community had created some standards, Microsoft had created an API that was used by some, but there weren't any broad standards. Many of these APIs didn't provide for scalability. In a call center environment there may be hundreds to thousands of calls at the same time. Scalability had to be taken into account. Anyone who has built Web servers knows this problem, but it wasn't understood so well in the telephony industry. The need for scalability ws something that those in the group understood from day 1.

Another need that came from back then was that the voice rendering had to be able to happen in the network. The rendering, both input and output, was very expensive at the time, and it did not run well on small devices. This model wasn't much like how the Web was working at the time, but it is very similar to how Web 2.0 applications that use AJAX, are working.

Another thing to remember is that the idea of basing this all on Web standards was a novel idea. At the time, telephony folks had large proprietary boxes that were installed in the phone network somewhere. We were very interested in bringing the Web architecture to this technology space. It was very important to build the new voice technologies on the Web architecture from the beginning.

It was a great selling point, we could talk with prospective customers, let them know that their existing backend infrastructure would work with a new voice interface.



Have largely solved bringing voice on the web to the telephony world.
Billions of phone calls each year browse the voice web with trillions of VXML web pages rendered to users. Our being driven by mobile device needs is a direct result of our success in the telephony world.

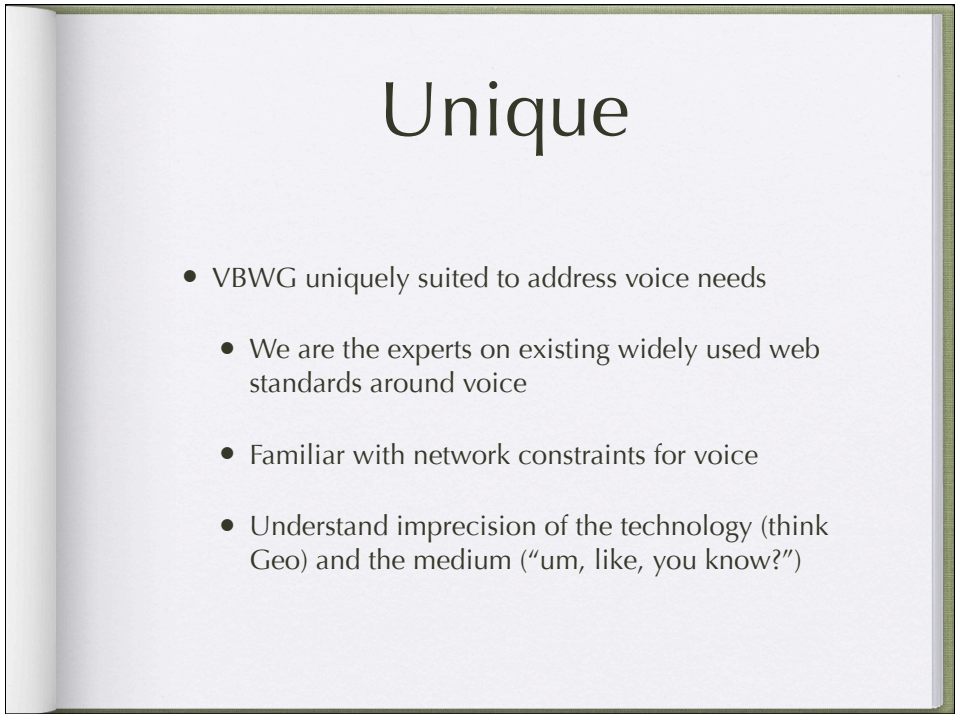
Today the VBWG is being driven very heavily by mobile voice needs.
All of the companies that provide VoiceXML technologies, they are all interested in mobile devices.
What does that mean for our group? It means we're paying a lot more attention to multimodal applications.
I hesitate to use that buzzword. Any application that has something you can look at, touch, hear, listens to you, is multimodal. It doesn't even have to be both visual and aural. Geolocation is a modality for instance.

In the VBWG we are very interested in small screen devices, or situations where peoples eyes or hands may be busy.
Today, we're very much driven by this.

Local context and search is something else we need to be integrated with.

There's also a new breed of developers drawn to the mobile space. Some of them may have never programmed for other devices before. It's important for us that we present this to them, even in simpler ways than are available today.

And yet, based on conversations we've had with others at TPAC, many think the group is focused purely on Call Center telephony.



To toot our own horn here: other groups often say it's better to create a new approach that throws out all the complex stuff our standards support and just do the simple stuff. But simple stuff just covers 50% of what you want. To get to 95% of what you want to do requires a whole heck of a lot of work.

There's a lot of expertise in the VBWG on voice technologies and the Web based standards for voice. We're particularly aware of the network constraints as well.

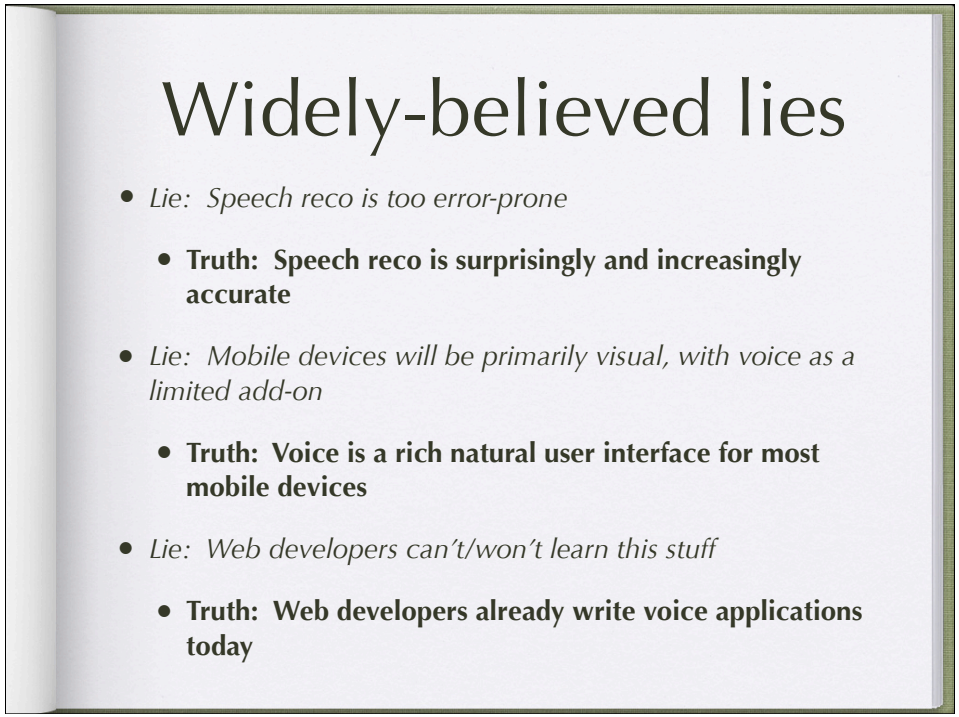
It's still important for us to support networked processing. Though the devices are getting more powerful the processing is getting more complicated.

Imprecision is also something Voice folks know well. I'm using imprecision in the scientific sense.
A good example of imprecision is geolocation. Whether it's wifi, cell tower or GPS based, there's an error factor to it, an imprecision to it.

The same is true from voice systems. There's an imprecision to it, that doesn't mean it's not useful information.

We have a lot of experience in knowing how to properly exploit confidence information when using an imprecise technology.

The voice medium is also imprecise -- when we talk we can mean many things, or we can say many words without meaning anything We have a lot of experience in how to not only deal with inputs given to the technology but how to encourage people to provide inputs that work better with the technology.



Going to talk about myths that surround ASR and TTS.
I use the word lies because after you leave this meeting, if you spread it, you're lying!
First up: speech is too error-prone.
Recognition built-in on small mobile devices is probably not as good as that in the network, but the state of the art has progressed a lot.
It's increasingly accurate, and as long as you are using it for the cases for which it was intended, it works fairly well for the majority of people.

The 2nd myth is that mobile device screens will be the primary input.
Originally the only mobile devices *were* phones, so clearly voice has been the primary input for a while.
We have a strong belief that voice will remain a strong component for a long time.

The last myth is about Web developers. We hear it's too complicated, etc. The best people to use this technology today is Web developers. I've worked at companies that both provide and use Voice technology.

A big plus has been to be able to tap into the Web development community.
Making the standards more accessible to the average Web developer is something we continue to do.

What are we working on today?
Three main high level goals: get out into the world what voice can do, what it does today in the call center world, what it does today in the mobile world, and what we expect it to do in the future of the Web.

We want to see future languages at w3c, including our own, play well together.

We haven't seen much integration between specs. This is something we've been giving a lot of attention to since the TPAC. There's been an increasing interest in this.

We're making quite a number of changes that generalize aspects of VoiceXML that are hard coded today. In VoiceXML 3 we're generalizing it in ways that allow developers to create their own paradigms for voice user interaction, and then allow other developers to make use of that paradigm. That will greatly simplify the coding for the latter group of developers.

We need you!
Don't just pass this along to your WGs (but do that too), we need you to do it. I've listed three sites with free access to developers.
They provide free hosting to try out Voice applications.
Each one has a tutorial for how to build applications.
Go build an app: build a front end to your phone, something that lets callers chose to leave a message, ask for a callback, whatever. It's easy to do.

Try it out. There are a lot of cool things you can do without a lot of work.

We need people on this call to work with us on building a combined group of people to build the best future applications. VBWG does not contain HTML experts.

The HTML WG probably does not contain the worlds' experts on voice either.

We think it'd be really helpful to get a subset of the two groups together to figure out how best to build these combined applications in the future.

Please, learn about voice, it's not that hard. And let's figure out how to create a combined group of experts for building future applications of the Web.