Hypertext to Hypervoice

Linking what we say to what we do

GEDDES

Author: Martin Geddes

Commissioned by HarQen Inc.

HarQen™ Telefonica

Hypervoice:

Linking what we say to what we do

Imagine a world where computers enrich our voices with superhuman powers; where voice is integrated into our social media just as text and images currently are; where our voice can be used as a communication tool at its full capacity: simple, powerful and rich. This is the world of **hypervoice**, where voice on the Web is as native and natural as hypertext.

If Web 2.0 was about transcending place, enabling us to become more interactive and social online through text and images, Web 3.0 will be about enriching conversation through the human voice. To secure the potential benefits, we must do more than just enable digitized speech; we must also re-think our basic purposes and the patterns of our voice communications so as to take advantage of the potential functionality of the Web.

Voice: gift or demand?

Our Western culture is dominated by visual media. We are surrounded by pulsating video screens, scrolling texts, and enormous advertisement hoardings. Nevertheless, the human voice remains a precious and powerful form of communication. It is the easy and spontaneous mode in which we account for our day, make important statements, and express our feelings and beliefs.

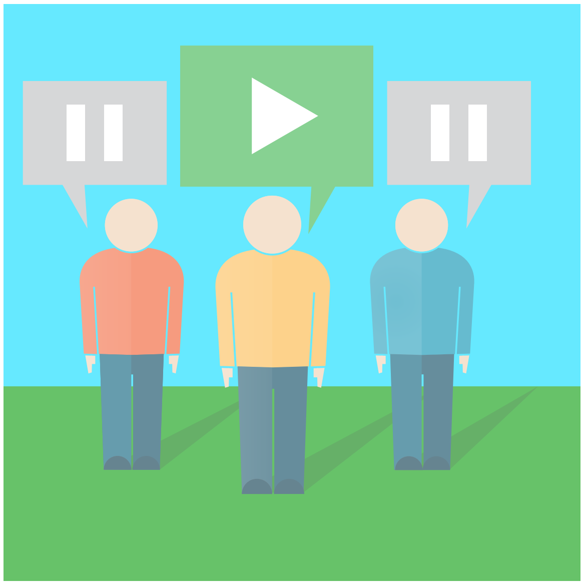
Until now, the free-form nature of voice communication has left it disconnected from the mainstream of our daily digital lives. Machines are adept at processing stored symbols, not spoken syllables. As a result, we have skewed our patterns of communication to favour media that can be easily structured and processed by machines.

Voice is casual and informal, well-suited to short interactions. Voice captures our creativity and so much else of our essence that makes us different from machines. Moments of inspiration or insight are often and easily communicated in this medium. Our most emotive and memorable conversations are vocal. Even Spielberg's E.T. didn't fax home; instead he found his voice.

However, there is a tension in voice between being creative and disorganized. The structured nature of non-voice communication gets in the way; it is friction to natural conversation. Nevertheless, we want to structure voice so it is amenable to the tools that amplify our effort, and makes it relevant to our personal and work lives. Too often, the use of voice becomes a demand, rather than a gift: it seems that there is too much cost involved to create with a structure, and yet without some helpful structure there is too little benefit.

Tempo and timing

A conversation calls for the participation of at least two people. As the number of parties to a conversation increases, so does the cost and complexity of making that interaction live and synchronous. Despite our valiant efforts on conference calls, the quality of the interaction appears to decline with additional participants.



Voice communication remains patterned on telephony, with its model of interruption. When A calls B, B doesn't know the reason for the call; the power in the conversation when B answers the call sits with A. Within this largely unexamined paradigm, 'live' voice is seen as the ideal, and voice-recording is presented as a managed failure mode that is bolted on as an afterthought. This leaves us with the dismal experiences of dictating or receiving voicemail messages.

Conference calls perpetuate these problems of rendezvous and synchronous communications. They are over-demanding of participants' time, given the customary practice of attending the entire call. Typically, though, many participants are barely present to the conversation: conference calls often run as background noise whilst executives look at their emails, giving the group conversation the status of a sort of corporate soap opera.

By contrast, we want voice to be alive, to feel alive, and to allow and encourage wide participation in conversations. Ideally the interested parties would be free of the obligation to be present to the original conversation. We want a fast tempo, and functional engagement, without the trouble of timing.

> *We want voice to be alive, to feel alive, and to allow and encourage wide participation in conversations.*

Inclusion and exclusion

In the past, the mix of computers and chatter has been unsatisfactory. Whereas computers are adept at manipulating stored information, voice is by its nature live and synchronous: we speak and await an immediate response. Historically, as a result of technological constraints, voice has also been ephemeral. Now that computers can readily make and store records of what we say, recorded voice as a feature allows us to increase the number of parties to a conversation in a way that is freed from the time constraints of 'needing to have been there'.

In the same way that hypertext did for text, hypervoice brings voice into the era of the Web.

Hitherto we lacked adequate means to index, search, filter and forward voice conversations. This has turned every recorded voice artifact into a liability rather than a digital asset, with the result that very few recorded calls are ever listened to or create value. Furthermore, our spoken words are often stuttering and provisional, with a high noise-to-signal ratio: we recognize that key information or decisions take up only a small part of the conversation. Yet our tools force us into an all-or-nothing approach to sharing voice. As a result, we often take on the burdens of transcribing the totality of a voice-recording into text, purely to record and share a few key parts of a conversation.

Voice is missing from hypermedia

There are three apparent conflicts around voice that we need to resolve:

- → How can we make voice better than using the structure of alternative non-voice media such as email or instant messaging?
- → How can we satisfy the need for both synchronous and asynchronous voice communications?
- → How can we resolve the desire for both privacy and appropriate sharing of recorded voice media?

These issues share a common underlying root: voice is missing from our concept of hypermedia.

We have thus far failed to understand and work with the natural dynamics of voice communications. Indeed, the Web was conceived without voice being a part of it:

> *HyperText is a way to link and access information of various kinds as a web of nodes in which the user can browse at will. Potentially, HyperText provides a single user-interface to many large classes of stored information such as reports, notes, data-bases, computer documentation and on-line systems help ... A program which provides access to the hypertext world we call a browser.*
>
> *– T. Berners-Lee, R. Cailliau,*
> *12 November 1990, CERN*

This tendency to unthinkingly place text in the ascendant was identified by Ted Nelson, who coined the word 'hypertext'. By 1992 he was pointing out (in *Literary Machines*) that the word 'hypertext' had become generally accepted for branching and responding text, but that the corresponding word 'hypermedia' – meaning complexes of branching and responding graphics, movies and sound (as well as text) – was much less used. "Instead they use the strange term 'interactive multimedia': this is four syllables longer, and does not express the idea of extending hypertext."

What's missing from the repertoire of hypermedia is hypervoice. In the same way that hypertext did for text, hypervoice brings voice into the era of the Web. However, Web technologies like WebRTC or VoIP are neither necessary nor sufficient to achieve this. It's the content that matters, not its delivery mechanism: how do you organize and flow value around voice?

Given the possibility of a richer use of voice contributions, we need to re-think our tools.

Adding voice to our activity streams

To progress we must reframe the challenge of 'voice', which in turn takes us back to the fundamentals of all human communications, in voice or any other medium.

Conversations are sequences of 'Gesture — Response', and these moves create an activity stream. These gestures and responses are much richer in their forms than merely spoken interactions and interjections, since we can also make parallel digital gestures —such as pulling up web pages, and advancing shared presentations. We make digital responses, such as taking notes after some else's spoken comment. Critically, these parallel activities are separated from their voice context, with the result that there is *no cross-relation between what is said and what is done*.

The management of our digital artifacts does not yet extend to live or recorded voice. Voice gestures aren't captured at all, or are captured as whole recorded conversations that have little value because they cannot be searched, navigated, or integrated into our workflow. And who has the time or the impetus to plough through an hour-long recording in real time, in search of a crucial but tiny snippet? To progress, we must create a *unified* activity stream, and enable richer conversations and the collaboration that can arise from a true integration of voice and text.

As we organize the gestures and responses that are our business conversations, we believe we have a choice of space travel versus time travel. With 'hypervoice' we transcend that limit: we can associate what is said with what is done using *when*, not *where*.

Given the possibility of a richer use of voice contributions, we need to re-think our tools. What does hypervoice mean in practice for the user experience?

> *Nobody needs a course to learn to use hypertext; the same is true for hypervoice.*

How hypervoice works

Hypervoice turns voice into a native Web object, rather than trying to convert it into a text object. It works by tying together all the gestures and interactions made during a conversation into a unified whole. The basis of the linking structure is the relative time in which gestures and responses occur.

When you make a note during an oral conversation, the words you type are connected to the words that have been spoken. Your typed note is important, and therefore forms relevant searchable context for the voice. You can tag a moment in the speech as easily as the piece of text you have typed. Indeed, *any and every* digital gesture can be tied to its moment in the spoken context. These hypervoice conversations can be given fine-grained permissions for sharing, searching and syndication, resolving the privacy issues around coarse-grained voice recording.

If you have an agenda for the conversation, this provides structure; as you advance through the agenda items, the relevant people can be included (if not already present), and those transitions are noted in the activity stream. Those agenda items can also be rich objects, tying the voice to its context. For instance, an agenda item might be a customer helpdesk request, drawn from a ticketing system: the voice segment on that trouble ticket can then be directly linked to the discussion about its resolution.

Thus hypervoice is a natural extension of the fundamental Web construct of linking, both to and from the conversation object, in a manner that is suitable for voice.

There is value to the user in creating metadata for hypervoice, as it is a natural byproduct of what people do anyway, and it transforms the voice into a digital asset. This in turn makes voice material easy to share, consume and distribute, and there is no learning curve, no barrier to utility, no friction. Nobody needs a course to learn to use hypertext; the same is true for hypervoice. By unbundling voice from synchronous telephony (or telephony-like unified communications services) voice, too, can become an 'anytime, anywhere' asynchronous medium.

No computer in our lifetimes will ever rival a human voice's capacity to conveying rich and complex social and emotional meaning.

Hypervoice is at the core of communications

Why put hypervoice as the central plank of your collaboration strategy? Because the conversations that really matter are conducted in the voice channel. No computer in our lifetimes will ever rival a human voice's capacity to conveying rich and complex social and emotional meaning. Hypervoice has the benefit of enriching voice's value, rather than trying to replace it with hypertext. It allows humans and machines to work with their strengths, rather than in opposition to them.

Just as hypertext breaks and re-connects the *place* metaphor for text, hypervoice separates and links the *time* elements of spoken conversation. As has become the case for all manner of text material, voice will no longer left to languish in the limited context of its origination.

Hypervoice allows other non-voice metadata and digital gestures and responses to provide the structure around voice, as a natural byproduct of our online behavior. By making voice easy to navigate using these structures, it makes voice recordings useful for those who are not present at the live event. Hypervoice makes voice suitable for sharing and syndication, because it can demarcate and identify parts of the conversation.

Hypervoice is like being hyper-present. It's 'better than being there' as it augments and enriches our conversations.

Hypervoice is like being hyper-intelligent. You can access everything ever said, with unbounded working memory.

Hypervoice is like being hyper-organised. You retain the link between recorded words or images and their spoken context.

The future of voice is hypervoice.

For more information on author Martin Geddes, visit: www.martingeddes.com

For more information on The Hypervoice Project, visit: www.hypervoice.org.