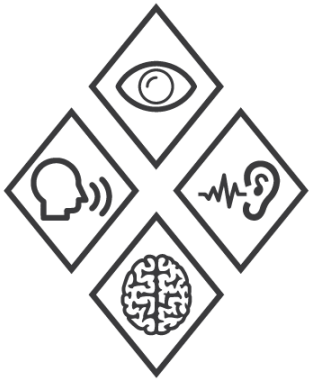
On Multiformat Communications



Alec Gordon
ag@beamplatform.com
2018

0. Intro

Text based communication, such as SMS, IP messaging, and emails have afforded us unprecedented flexibility to live our lives while remaining hyperconnected. Much trivial information gets exchanged this way, interspersed with the occasional bit of useful data one wishes to keep – a name, an address, a date, and so forth. When used for work, such tools are indispensable due to the sheer volume of transacted info, asking us to keep track of people, events, processes and objectives. While spoken language remains the fastest method of conveying information, written (typed) language is the one that stands preserved throughout time. We know full well that records exist of our digital conversation for speedy search and retrieval. Our "surrogate memory" forever at reach, freeing our mind to focus on higher order functions. With that simple comfort, we remember to forget as we navigate the increasingly noisy world. When a friend begins to tell you a new address or number, an automatic reply follows – "Just text it to me".

With capable computers in every pocket, we opt to use them for most tasks in favor of other, less portable devices. For rapid exchange of information, text messaging has proven to be the optimal choice. Aside from the digitized nature of record, the messaging interface is both intuitive and clean. Left/right author segmentation shows turns taken with typing/reading, while metadata such as timestamps provide a clear flow from one correspondence to the next. A conversation, whether by text, in person or over a phone follows similar back-and-forth structure, and yet, only one is autoformatted and preserved on the device. Of all digital means of communication, the phonecall remains the last medium that has yet to be properly "digitized". Yet this is where we're headed.

A simple notion presents itself – apply to phone calls the same interface employed by messaging technologies, with simple left/right segmentation, sequence of events and metadata, and put it in everyone's phone. Conduct phone calls effortlessly, without the need to jump to a notebook whenever something needs to be remembered, knowing that a record can be accessed with few taps of the finger, with no penalty to us. What freedoms would that afford? Our memory, supplemented with yet more computing power and analytical prowess, exists to tackle bigger problems and make associations for which there is no substitute.

The last decade gave rise to many of the critical components such a system would inadvertently use, yet they remain siloed. Sound capture and call recording provide audio files that are neither indexed nor searchable. Speech recognition, as a post processing method, outputs a transcript that isn't formatted nor are the words in text linked to audio. Typed messages aren't easily read aloud without the use of 3rd party applications. And since this isn't done in real-time, much of the utility of such tools goes unrecognized.

We propose a framework for the first true Multiformat Communications platform converging speech and text into the same unified interface. You talk on the phone, and real-time text is being sent to the other device simultaneously with your voice, and vice versa. What's spoken can be read, and what's typed can be heard. A word typed and a word spoken are now two sides of the same coin, transmitted in sync with one another, opening up voice communication to the same technologies that let us parse, index and search strings of text. An interface described herein is optimally suited for such tasks.

Contents

1. Unified Communications 2.0

The phone of tomorrow is bound to offer new levels of interpersonal connectivity with more novel functionality and subsequent utility. With custom software being wedged into more and more places, surely our networked interactions stand to benefit greatly from such advances. The nearest technology to take on the role, Unified Communications, has been in existence since the 80’s. While the promise of unifying chat, voice, presence, video and web, along with data sharing has been enticing, the promise largely fell flat. This concept was not necessarily a single product but a suite of technologies mashed together, linked under a common banner.

Focusing on two key methods of interaction, real-time text and calls, we set out to redefine what communication looks like when truly unified. The goal was to make them interchangeable – being able to talk by text messaging, then, walking off, seamlessly switch over to voice, all in the same maintained session. Ability to do so meant carrying over the record from one medium – text, to the next – speech, and standardizing both into an uninterrupted thread, accessible by a single interface. Multiformat transmission of information, with multiplex data being carried over concurrent streams reuniting on the receiving device, solved this problem. Coupled transfer of this type already exists, such as in case of Telecommunications Relay Service (TRS), which happens to have a human intermediary performing the “conversion” from one format to the next. Multiformat transmission supersedes that in design, as both formats are relayed in sync, without the need for human intervention.

A key component of Multiformat technology is the conversion from speech to text. With the proliferation of speech enabled devices and hands-free interfaces, people have grown accustomed to scheduling calendar events, creating shopping lists with one-off commands. To say speech recognition has evolved would be an understatement, and yet most deemed it too unreliable to be useful in the medium of phone conversation. 5.1% is the magic number to beat – this is the accepted human word error rate, while exceeding that would make the software, by nature, “superhuman”. Today we stand on the cusp on that threshold, as 2017 brought us two breakthroughs from tech giants IBM and Microsoft, first homing in at 5.5% error rate [1], and latter matching the 5.1% “human parity” mark mere months ago [2].

Extending the messaging UI (historically asynchronous and turn-based) to speech (synchronous and real-time) require that we adapt our messengers to also work in true real-time, allowing the streaming of text as it’s typed, rather than when a “Send” command is executed. This was first demonstrated with Beam Messenger, and it received global media coverage for it’s novel method of preserving 1-to-1 chronology of conversation. When rendering a phone call in text, this method proved to be the optimal solution. In future, our messaging systems will be adaptive, switching from turn-based to real-time as appropriate to the conversation, and real-time metadata will be used for more than just chat, but things like user authentication, mass pattern recognition and beyond. Considering the recent standardization of real-time text, the vast majority of use cases for this promising new technology remain to be uncovered.

2. The road to “digitized self”

On the quest to self preservation, people have adapted their lives based around digital tools that promise to capture the fleeting moments, turning them into data. For everything we capture – notes, pictures, tracked steps – we easily dismiss the largest set, our spoken language. But we must start… To think one day with revolutionary new sensors, by simply having a thought we could interface with powerful embedded computers, and navigate our environment with superhuman precision. Or with the coming singularity, we can extend the “self” onto machines, transcending our fraying bodies onto the digital realm. Without data, this will remain the stuff of fiction, and the sooner we begin to capture it, the more complete the end result will be.

We’ve come a long way since Vannevar Bush first posited such visions in the quintessential 1945 essay “As We May Think”. What we now call “the cloud” he called “the memex” [3]. Such a contraption would fortify one’s memory with permanence, and navigating it would closely resemble that done by associative thinking – not unlike as we may think. But simple data capture is not enough. A record must be stored, extended and easily consulted, as too much data gets unwieldy without proper indexing tools. In current age, to simply capture audio without persistent hyperlinked transcription therefore is wasteful and inefficient.

Due to the volume, speech is most dense human form of signal transfer, and thus imperative to constructing a digital data bank about a person. A model created using this set of data would trump most other models, as those looking at social media posts, smartphone pictures and the like, are less relevant to the inner workings of the mind.

Due to the interchangeable nature of Multiformat Communications, events from separate conversations, normalized to a single format, can be strung together and analyzed.

The notion of digitally creating a “self” was further explored by Gordon Bell during his MyLifeBits lifelogging project while at Microsoft. Goal was to collate all the external data one could justifiably get, utilizing a variety of hardware sensors and software programs, and string them together in a SQL database for total recall [4]. Key part was the addition of metadata input, which led to the feeling of having an all-powerful “surrogate memory”, freeing you to greater work. Such automatic capture, as in the case of Multiformat phone calls, allows more correlations to be made between abundant info. Bell further stated that “users are not just unwilling to classify, but are in fact unable to do it”, and that special know-how must be employed to make the data useful. For all the info one could get from digital communications, therefore, it is important that a single format, real-time text, is maintained when switching between chat, call, video, BCI, and beyond.

Supplying developers speech data through such real-time interfaces allows for the next wave of autonomous and predictive technologies to emerge. Far beyond memory reinforcement, one can even imagine analyzing spoken language for signs of cognitive decline, such as in dementia and Alzheimer’s, through computational linguistics [5]. Diagnostic and monitoring efforts will rely on speech patterns more. The quicker we transition to this new paradigm, the more complete the picture becomes.

3. 5G, 5G everywhere

With 5th generation mobile networking on the horizon, and early deployments slated for as soon as 2018 [6], we must pause and think of what is ahead. Surely our download speeds are bound to impress, but what is beyond, what is truly novel? What new opportunities will emerge when data throughput shoots skyward from 0.1 Gigabit/s to 10 Gbit/s, and latency shrinks down from 100 milliseconds to ultra low 1ms? With 100x leaps throughout, we are approaching territory when our devices are no longer bound by their internals, and the mystical “cloud” starts to feel a lot more familiar, and “local”.

This notion of “local cloud” extends beyond simply maintaining a large repository of data over the Internet, but rather that there isn’t a distinction between accessing content stored on the device, and content stored remotely. With single digit latency, streaming of media feels instantaneous. Smartphone storage becomes a thing of the past, as ubiquitous connectivity coupled with remote processing pushes us further into the realm of invisible computing.

This notion bodes well for Multiformat Communications, and future proofs it for the next wave of telephony. An explosion of data is bound to be seen, with simultaneous archiving of speech, video and text, all from a single input. When used optimally, Multiformat allows one to switch between the varying ways of relaying a message -- typing one second, speaking the next -- and showing right after. The text provides cohesion and continuity between all modes, with lightweight annotation of digital interaction.

Speech and audio, in turn, take up considerable space, and hi-res video yet more. For this to behave the way we need, storage and streaming of media files will take place off-premise, kept at the ready for immediate viewership.

Other key services will tap into 5G as time progresses, and software is built anew around the fresh paradigm. Digitally conversing in groups, whether including 3 people or 30, allows all to interpret the same events without delay, regardless of the interface each person chooses to communicate – typing, speaking, etc. One could envision a future whereby 4 people all talk in real-time by way of real-time text, one by way of voice, another by way of telepresence, and the last by way of a “talking head”, a.k.a. 3D facial animation from audio/text input [7]. Tying them together is a massive amount of codec work, all done remotely and streamed within milliseconds to each participant. Now throw a foreign language into the mix, and you can see, the future has got to get quicker.

4. Everyone’s personal assistant

The chatbot revolution has taken the world by storm. The messaging UI has proven to be a capable interface, allowing people to interact with machines in a way most natural to them – through conversation. Although text currently dominates the recent surgence of bots, from weather bots to pizza bots, the laborious nature of typing has quickly shown to be suboptimal in the fast paced world of already speedy touch interfaces.

Voice quickly is encroaching on our homes, with the proliferation of smart speakers, smart TVs, and smart toasters just over the horizon. The next logical step is integration of similar technology within our telecommunications. While texting a calendar entry to a bot is viable in theory, extracting that intent from free-flowing phone call speech is the next natural step towards phone 2.0.

In the world of tomorrow placing a “call” to your digital agent will feel no more unnatural than ringing a human assistant today. Omitting the niceties, you quickly pace through the laundry list of commands – “schedule lunch”, “check traffic” – hear back a confirmation, and on you go. Unlike asynchronous texting, where people can wait, with bots you want a rapid back-and-forth. For this type of scenario, the speech interface is unparalleled. With Multiformat Communications however, this interaction is already possible as speech and text converge, and you’re able to access the countless text-only bots already in existence. This “silent call” has only the human speaker vocalizing, bypassing the tired “initiator command” each time, with chatbots typing back in real-time. As reading is faster than hearing, this interaction is optimized for efficiency. If line of sight cannot be maintained, the bot begins to synthesize speech.

As phone calls transition to Multiformat, a new category of digital agents will emerge. The first kind, “Passive Personal” (PP), exists assigned to every user. Only you can interact with it, and only you can hear it & read it. Your personal “SAM”. When info is exchanged pertaining to one person, as when asked whether “you are free next Thursday after work”, PP bot concisely types out or speaks out that “you are busy/available”. This agent is reactive, but can be also be invoked with a spoken command, responding to the voiceprint of the assigned user. When a call ends, the user maintains connection to PP bot, and may continue to issue commands, notes, while the mind is fresh.

The second kind, “Active Mutual” (AM), exists between you and the person/group you are speaking to. When AM bot types, all sides can read the message, and when AM bot speaks, every user can hear it and interact with it at will. This agent is active, and is particularly suited for suggestions and recommendations, as well as maintaining group lists, etc. Each and every continued conversation has its own AM bot, distinct from any other AM bot. Two AM bots cannot interface with one another, however your PP bot can interface with any group AM bot you’re party to.

Due to the nature of Multiformat Communications, and the preservation of persistent records regardless of method – speech, text – these digital agents have volumes more training data and are better attuned to your profile.

5. A “search engine” for your brain

Peering into one’s own mind with newfound capabilities of AI is no longer out of reach. Providing computers with new relevant training data allows them to learn us better, and extract sentiment where none was possible before. Standardizing all our communications to a single unified format, with technology described herein, opens the data to draw meaningful connections more often. An interface for a truly “personal machine”; a “memex”.

With phone calls now searchable, every instance of a spoken name, address, or date is easily highlighted and added to special “entity library”. Beyond simple lookup, new cases emerge with the ability to analyze text data across the full spectrum of conversation, as there are no gaps between talking and typing. Every case where the name “Ray” was spoken can easily be served up. With rising sophistication in AI, NLP and audio spectrum analysis, more abstract questions can be asked, such as “am I too loud?”, “when did I first hear of this?” or “do I make too many promises?”.

Indexing has always been the bigger problem than compression. To fully tap into the new data, more powerful tools must be employed. While rudimentary text-to-text search is the first tool to consider, by aligning text + audio in real-time, both formats can now be utilized when searching for keywords. The first step is speech recognition, and the second step is in creating a spectrogram out of the audio file, and training Deep Neural Networks (DNN) to make out the contents, producing a searchable index that can be easily queried by either speaking a keyword or typing it out. This gets around the inherent imperfections of speech recognition whereby it mistranscribes the words, thus rendering any text-for-text search useless.

A number of such “DNN for sound” tools have emerged [8], ensuring the highest degree of search accuracy throughout.

“Part of feeling secure is knowing that capture is automatic.” Considering we’re already using computers to conduct phone calls, the logical extension of that is to use the newfound power to start building up a personal “databank” of stated facts, personal knowledge and memories. A more intuitive way to recall is by placing “voice bookmarks” in the moment information is transacted. Devoid of the minutiae of notetaking, we are in turn freed to shift our attention towards more meaningful and rewarding deep work.

Our brains are “association machines”, weaving together memories as trails of itemized facts. When one fails to recall, the memory can often be jogged by a nearby factoid, getting you closer to the destination. In the case of Multiformat, the full spectrum of conversation is preserved – important bits as well as the trivialities. Not to be discarded, it is often those trivial facts that help remember the connection. The more we capture, the more of our memory we can offload to machines. They simply don’t forget.

Considering the ethics of such auto capture, the notion of the Big Brother comes to mind. Consider this – the Big Brother gathers data at no recourse, without offering any utility to us. We produce, yet we get nothing. With Multiformat calls, you own it all. The data is at your fingertips, serving and helping you the moment when you need it.

6. The everyday polyglot

We are fast approaching the oft-cited Universal Translator, and it has become easier than ever to break through language barriers. Ubiquitous web services allow us to translate text, and now we’re seeing similar tools become integrated into our communication services.

The task is rather simple – adopt a universal language, or use technology to mend together ones already in use. In 1930’s, early work by Bell Labs demonstrated an analog speech analyzer (the Vocoder), and subsequent speech synthesizer (the Voder). An addition was postulated to combine the Vocoder with a stenotype, yielding speech recognition, and further yet, to combine it with a translation mechanism for end-to-end universal translation. Between the two avenues, the gap is closing faster on the latter.

In recent years, Microsoft leapt forward in this quest with the introduction of Skype Translator, to date accounting for over 10 spoken languages and even more if typed [9]. With even headphones getting smarter, the software based tools will offer more versatility and appear in more places. In time, we will outgrow our reliance on pocket computers, with IoT sensors picking up the slack, immersing us and bringing us closer together. As the best technology is also invisible, the thought of being universally understood may not be as far away.

The Multiformat Communications interface, deployed across many applications, will therefore democratize this common “superpower”. Many second and third language users will find new freedoms, unburdened to communicate the way they choose.

This level of flexibility is bound to uncover new possibilities for inter collaboration, at once bringing us closer and turning us more efficient.

As every past and present Multiformat call is also encoded with synchronized text, it is possible to view this text in other chosen languages also. Translation is done automatically, and cycling through languages to view a sentence or paragraph becomes seamless and second nature. For those multilingual among us, the limitation of using only a single language to communicate is gone, as we may pick and choose words & phrases that best describe the situation, while letting machine translation handle the rest.

Searching for terms is thus improved also, unbound by the origin language of the queried text. As people begin to communicate more freely - switching between dialects to better express themselves, the ability to find any keywords and terms remains frictionless.

We’ve come a long way since the Voder first broke new ground. In present day, applying Deep Learning models to as little as 20 minutes of target’s speech, new software is able to produce any sound in target’s voice, whether they spoke it or not [10]. When combined with Multiformat texts and call, this remarkable technology will let one type, and the other to hear the words spoken in the very voice of the originator. Any translation thus evolves beyond simple text, and into territory where anyone, even you, can say a word a hundred different ways.

7. Universal design

Universal or Accessible design, among other things, involves putting the human first in any Human-Computer Interaction paradigm that is being constructed. The human is fallible, unreliable and inconsistent, the computer – none of those things. It is therefore easy, even tempting, to design models with all the perfect assumptions, and end up with a largely unusable system that fails to connect with people, or worse, exceeds their full capabilities.

In using technology we take much for granted. Our interfaces and applications have been fine tuned to work intuitively, while applying as little cognitive load as possible. We tend to appreciate them more when these technologies break, and even more so, when we ourselves start breaking down... “Having a surrogate memory creates a freeing, uplifting, and secure feeling – similar to having an assistant with a perfect memory” – and we will welcome these assistants as our own memories get frayed with age.

Universal Multiformat design ensures preservation of fact and knowledge, all that was communicated through speech and text. Making the interface to data accessible through every smartphone, computer and device, is simply good practice in the age of info overload and attention strain. Relying on multiple ways to “read” the data carries further benefits. One must not be stripped of a sense in order to use another. While the hearing impaired would stand to benefit greatly from a Multiformat call, so would those in loud settings.

Those who are speech impaired may be unable to respond vocally, opting to type the message instead, but so might one in a room that’s quiet.

The playing field is leveled through the use of new text-to-speech technologies, creating an environment where either option to type in real-time producing your own voice, or speaking, are equally compatible with one another. A driver will have no more trouble “texting back” in a group chat convo than if he were using his hands to type. He hears, and they read.

At the core of Multiformat Communications is real-time text technology – the key component for combining synchronous speech with (now) synchronous text. And it is here to stay. As the December 2016 ruling by the Federal Communications Commission shown, real-time text is simply the most accessible and inclusive form of messaging around. It is henceforth required on every modern phone, and is the new telecommunications standard [11].

The phone of tomorrow will offer unprecedented flexibility in how we choose to communicate, relate and share. It starts with incorporating the principles outlined herein, and extending the functionality as we move forward. A simple interface allows a new spectrum of future possibilities and use cases to emerge, and it is up to us to build that future.

8. References

1. https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/
2. https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/
3. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
4. https://www.microsoft.com/en-us/research/project/mylifebits/
5. Fraser K, Meltzer J, Rudzicz F (2015) Linguistic Features Identify Alzheimer’s Disease in Narrative Speech. Journal of Alzheimer’s Disease 49, 407-422.
6. http://www.multichannel.com/news/finance/verizon-exec-meaningful-5g-deployments-start-2018/411354
7. http://research.nvidia.com/publication/2017-07_Audio-Driven-Facial-Animation
8. http://blog.deepgram.com/introducing-deepgram-brain/
9. https://blogs.skype.com/news/2017/04/06/skype-translator-offers-japanese-10th-real-time-spoken-language/
10. https://blogs.adobe.com/conversations/2016/11/lets-get-experimental-behind-the-adobe-max-sneaks.html
11. https://apps.fcc.gov/edocs_public/attachmatch/DOC-342624A1.pdf