STRATEGIC NEWS SERVICE®

GLOBAL REPORT ON
TECHNOLOGY AND THE ECONOMY

Publisher's Note: If there is one technology in the world which would seem to hold the key to liberating the human-compute connection, it is voice recognition; and if there is one person in the world who knows most about this link, it is likely Vlad Sejnoha, CTO at Nuance. Most of us have spent a lifetime waiting for this technology to reach maturity, but Vlad is a star in the world of those who have made it happen.

But as SNS members will see from our issue this week, Vlad's interest in this technology goes far beyond VR and into what we have started calling "meta biometrics." I doubt that there is a major technology provider today that is not interested in getting this into its products in one way or another, nor is there a major consumer or enterprise company that is not on the hunt for how this can be used to improve its own customer relations – and profits.

As Vlad's efforts lead to ever-greater understanding of man-machine communications, and of the information content in our own voices, the value of this work climbs exponentially. – mra.

MOBILE VOICE AND NATURAL LANGUAGE UNDERSTANDING: REVOLUTION OR EVOLUTION?

by Vlad Sejnoha, CTO, Nuance Communications

In 2012, mobile voice reached a critical tipping point. Voice technologies – once treated as fanciful science fiction in TV series like *Star Trek* and *Knight Rider* – began to be integrated into user interfaces for some of today's most popular consumer products. Companies such as Ford, Samsung, and of course, Apple are promoting voice technologies as key differentiators in primetime advertising, generating unprecedented consumer awareness and demand.

This momentum is being driven by acceleration in several mutually reinforcing business and technology catalysts, including natural language understanding (NLU) and artificial intelligence (AI) – a powerful combination that has made the mobile user interface (UI) smarter and transformed it into a far more natural way to communicate.

Advancements in speech, NLU, and AI will continue to transform the mobile user interface even further in the coming year. In fact, looking at 2013 and beyond, the mobile UI will eventually make interacting with mobile devices as natural as communicating with other people.

Hence, the tipping point. Some observers – commenting on the launch of Siri, the Apple iPhone's virtual assistant – have described the current proliferation of voice and natural language processing into the UI as the "third revolution" in computing, following the introduction of the graphical UI controlled by a mouse and the touch screen as the first and second, respectively.

But, as those in the voice technology industry well know, voice and NLU applications are hardly new. Is it really a revolution, or rather a fast evolution, bolstered by the marketing muscle of the handset manufacturers looking to differentiate their products?

The progress seen in voice technology in just a couple of decades is remarkable. Early systems required laborious training by users and could process only a few dozen words spoken in isolation (and not very accurately at that). Today, conversational mobile interfaces are able to understand natural user input by invoking nearly instantaneous, accurate, cloud-based, speaker-independent, large-vocabulary voice recognition and natural language understanding. Users expect their mobile devices to convert utterances such as "Where can I get some great sushi near me?" to text, infer the meaning, and retrieve the appropriate information from a cloud-based service.

The Catalysts for Voice

At the root of this progress is the continuous improvement of computing power, which enables the embedding of ever more sophisticated voice technology on mobile devices and fuels the growth of cloud computing. In addition, the rapid spread of high-speed networks allows speech application providers to instantaneously access larger and more accurate server-based engines in a way that's invisible to the user.

Handset manufacturers are also driving the adoption of voice technology as a means to address the challenges created by the homogenization of handset operating systems. These OEMs [original equipment manufacturers] are keen adopters of any technology that promises to enhance and differentiate the user experience on their particular devices.

Exponential growth in both the number of new apps for smart mobile devices and the amount of available content are stimulating mobile consumers' boundless appetites for new devices and innovations.

The demand for content and services on mobile devices has amplified the need to improve the usability of the prevalent mobile UIs. Often likened to "shrunken desktop metaphors," these UIs stymie users by forcing them to search through multiple screens to find apps, navigate deep menus, and scroll through lists of links. Voice input, coupled with natural language processing, promises to offer a way to cut through all the clutter.

All of this is happening at a time when voice technology has attained unprecedented levels of accuracy. Earlier generations of voice applications have accumulated vast quantities of user data – the lifeblood of statistical models. Those earlier generations of voice recognition technology relied on hand-crafted, restrictive grammars, but these are increasingly giving way to automatically trained statistical language models (SLMs). The SLMs lower the development burden, but also more readily support free-form user input.

Voice recognition systems are therefore becoming much easier for non-experts to work with, and the availability of inexpensive, cloud-based speech application programming interfaces (APIs) have attracted huge numbers of developers who are now experimenting with novel ways of adding voice to mobile applications.

Natural language processing itself has evolved, as well. Relying increasingly on powerful combinations of linguistic frameworks and data-driven machine learning, NLU is smarter and able to do more, including performing higher-level reasoning to complete tasks.

The great majority of voice and NLU systems are now being accessed from devices with keyboards and touchscreens, enabling application designers to make use of a much broader range of input and output modalities than in the past, including type, touch, trace, and handwriting. These not only add flexibility and convenience, but also provide a safety net that allows app designers to be bolder than they could be in the era of voice-only "interactive voice response" (IVR) telephony systems.

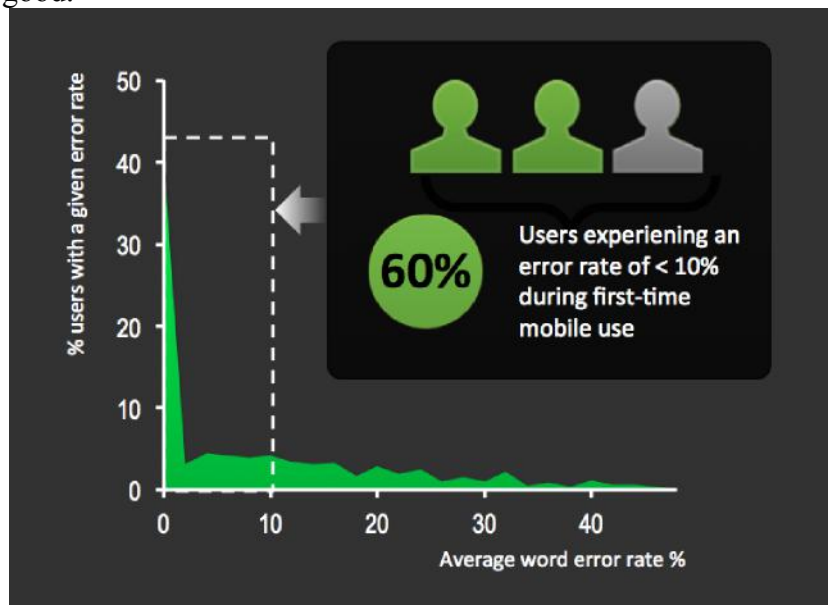
Improved connectivity has created a "network effect," making it possible not only to reach powerful off-board voice recognition and NLU instantaneously, but also to transparently integrate cloud-based services – including question-answering systems – that make voice UIs that much smarter and more valuable.

Finally, mobile voice and NLU are now getting far greater exposure to the public than ever before thanks to unprecedented advertising budgets and placements, such as Apple's Siri advertisements featuring A-list celebrities and, even more recently, Peyton Manning showcasing Buick's IntelliLink capabilities. Samsung advertised its Smart TV ("Hi TV") during the 2012 Academy Awards, for which the cost of a 30-second TV spot can equal what not so long ago might have been the annual research budget for a voice technology startup.

Measuring How Far Voice Recognition Has Come

How good has voice recognition become? The most common way to measure this is to look at the average error rate of a user population. However, it can be more revealing to look at the distribution of user error rates, which shows what fraction of the user population experiences performance that has surpassed a usability threshold.

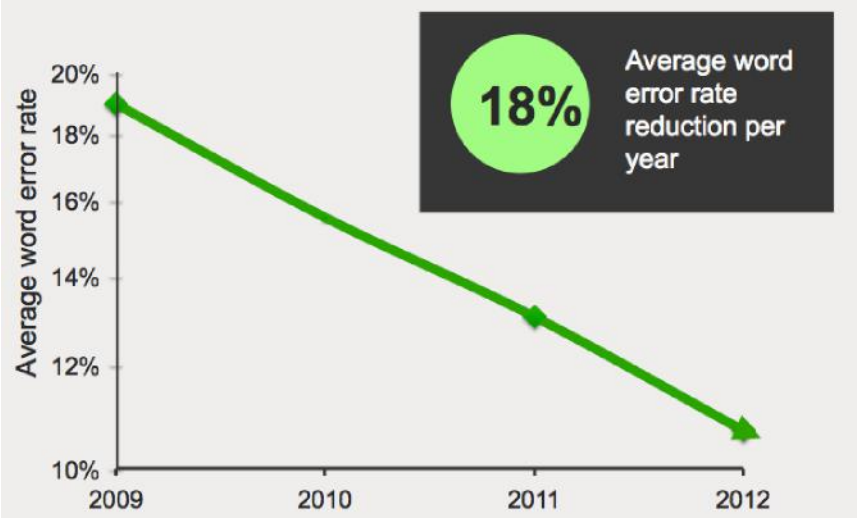
The chart below reflects the current state of the art. It shows the error rates for a sample of 10,000 users of Nuance's cloud-based speech API across a broad range of environments and a variety of handsets without any adaptation. As depicted, 60% of those users experience an error rate of 10% or less out of the box – even before the system has had a chance to begin to learn about the individual user. This is getting very good.



Source: Nuance Communications

But note the "'tail" of the distribution – the fraction of the user population whose speech is still a challenge for our systems, whether because it's disfluent or strongly accented, or the environment is too noisy.

The tail remains the focus of intense research, and the chart below shows how we've improved over time – basically, there has been a very consistent ~18% reduction in the average error rate per year. This means that more and more of the distribution in the first curve will be moving left, and more people will get superlative performance straight out of the box. This is critical to both application adoption and user retention.



Source: Nuance Communications

The improvements arise from a range of factors. Over the past five years or so, the amount of data Nuance uses to train its models has increased a hundredfold, and the models themselves are far more sophisticated.

Recently, for example, there has been a resurgence of the use of neural networks for voice recognition in the form of "deep belief networks" (DBNs). DBNs are loosely based on theories of how the brain performs pattern recognition, and while they still present substantial challenges – the training is very computationally expensive, and ensuring convergence can be difficult – compared with conventional techniques they are more effective in taking advantage of the available data and in minimizing the error rate.

Amid the continued improvements, Nuance is simultaneously making the tasks more difficult as voice recognition is further introduced into cars, livingrooms, and other acoustically challenging environments.

The Year of the Natural Language Understanding Application

The advent of high-quality plug-and-play cloud voice recognition has led to a plethora of new third-party voice applications. By late 2012, the number of participants in Nuance's developer program for cloud-based voice APIs had surpassed 13,000, with hundreds of apps being released via app stores. Google and AT&T offer similar programs. This vigorous uptake drives the exploration of new application concepts, faster update cycles via app stores, and learning from user feedback and data.

One of the fastest-growing apps is the voice-enabled "virtual assistant," which was popularized with Apple's launch of Siri for the iPhone. Many believe that these assistants will play a key role in alleviating a problem associated with the advance of technology. Although intended to make our lives easier, technology also has created data overload – an environment of constant interruptions by messages and flows of information. As voice-enabled assistants become smarter and more humanlike in their interactions, they could become a key tool for helping individuals cope with data overload. The availability of developer voice APIs has led to an explosion in the

number of "Siri clones," which perform commands, launch applications, conduct semantic search, and engage in chat.

Some are impressive, others less so. Objectively characterizing the quality of an NLU application can be challenging. Contrast this with voice-to-text, where arguably there is one primary concern: how accurately does it perform the conversion?

In NLU, the understanding accuracy of input concepts is clearly important. However, there are other important considerations, including:

Robustness to out-of-domain inputs: Does the system fall apart, or does it recover gracefully?

Efficiency: How effective is it in performing a task – how many steps are needed?

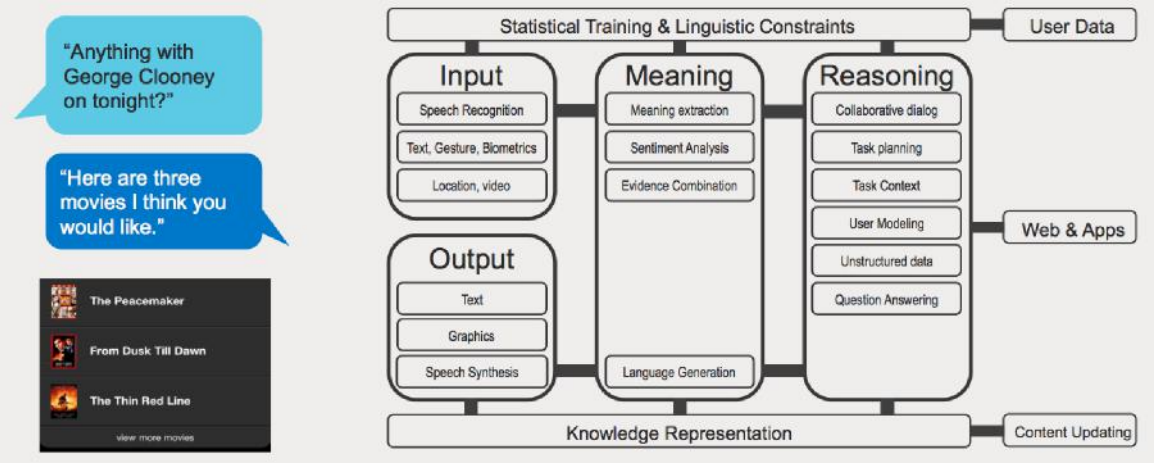
Breadth: How many functions, applications, and domains can it handle?

Depth: How intelligent is it within each domain – how many concepts and relationships can it tease out?

Character: How appealing is its character, if it has one? Should it have one?

Smarter NLU... or AI?

The better systems go well beyond the traditional voice-understanding architecture, which featured voice recognition; limited, grammar-based meaning extraction; and fixed prompted dialogs, typically in voice-only telephony implementations. The chart below shows the main constituents of a modern Voice/NLU framework.



Source: Nuance Communications

On the input side, modern NLU can handle information conveyed through a variety of modalities, gestures, and metadata, such as location and time of day. Extracting the meaning has become more sophisticated – today's NLU is able to infer higher levels of meaning, such as the user's sentiment; combine evidence from a variety of sources; and estimate its own confidence in the results.

Free-form conversational input demands dynamic, mixed-initiative dialog control. Gone are the scripted flows in which every possible path must be pre-defined. We now use efficient specifications – for example, functional combinations of reusable dialog agents that specify the desired high-level goals of an interaction but allow freedom in the manner of reaching those goals.

Sophisticated conversation management thus dynamically determines how to best advance an interaction toward a specified task goal, and is beginning to blur the line with "reasoning," which is squarely the province of Artificial Intelligence. Such reasoning is performed on high-level task representations that are increasingly dynamic – allowing the system to effectively program itself along the way. The choices the system makes are informed by explicit representation of user preferences, the interaction history, and the task context.

Whereas earlier AI systems proved to be brittle, today's state-of-the-art systems rely on more flexible and robust approaches that do well in the face of ambiguity and produce approximate solutions in situations where an exact answer might not be possible.

Modern NLU also includes novel functionality, such as question-answering from unstructured data, as exemplified by IBM's *Jeopardy!*-playing DeepQA project, Watson.
On the output side, these systems communicate with the user through a variety of means – visually, through text, and of course, by voice. Text-to-speech itself is increasingly natural sounding, with emotion, expression and inflection – in part because it, too, is being informed by an understanding of the underlying concepts in the text that is being synthesized. In other words, the systems understand the text they're reading aloud and automatically determine the right tone and emotion to apply.

NLU apps must understand increasingly large numbers of concepts, with complex relationships in many domains, and make use of knowledge representations known as *ontologies*: structures that capture the concepts and relationships in a reusable and maintainable form. Ontologies are used by the understanding components to extract concepts from spoken input and to determine what attributes to look for.

Probably the biggest change is that many of the models used by today's best-understanding systems blend rule-based and statistical approaches to take advantage of their respective strengths and minimize their weaknesses. Rule-based (or symbolic) models can efficiently and quickly encode repeatable and robust linguistic phenomena. They are increasingly being complemented by machine-learning algorithms which are very good at discovering patterns from data. Statistical NLU can learn to automatically find new "named entities," such as city names in familiar contexts, or discover new contexts for familiar entities and thus effectively "populate" the linguistic frameworks.

NLU Means Direct Access

One of the most powerful benefits of NLU in the user interface is its ability to cut through the layers of the visual hierarchy: multiple screens of app icons, folders and subfolders, or a sequence of intermediate Web pages. By simply speaking, a user can go directly from "intent" to the desired destination or action in one step.

By performing a "semantic" search, applications such as Dragon Go!, T-Mobile's Genius Search for its myTouch series, Dragon Assistant for Intel-inspired Ultrabooks,

and Dragon Mobile Assistant take users directly to useful destinations for a large proportion of high-frequency user search queries. By understanding that a query is about a restaurant reservation, for example, the system can navigate to, and populate the reservation form on, the OpenTable website, rather than displaying a page of blue links.

This is not an elimination of search as we know it, but a powerful complement that is quite transformative – a lot will now be reachable through only one utterance. While the search providers are beginning to experiment with replacing the links with direct answers for some queries, they are not allowing users to skip directly to an intended destination, omitting the intermediate steps. The "disintermediation" of the search portals for many query types may prove quite disruptive to the search industry.

How Do We Want to Talk to Our Devices?

All of this new functionality begs the question of how to best incorporate natural language understanding into today's visual interfaces. There is a broad range of possible approaches, including the "virtual assistant" category, which had a banner year in 2012 with Siri, Dragon Go, Dragon Mobile Assistant, Google, and nearly 60 others on the market.

When comparing virtual assistants or natural language understanding applications, the accuracy of the speech recognition itself is just one of many questions that must be asked. How robust is the NLU to inputs that were not anticipated by the designers? Does it fall apart? Does it behave intelligently? How efficient is it? In attempting to achieve a goal, how many steps are required? How many functions does it support? How deep is its understanding within any one of those areas? How portable is it? Does it have a persona? And if so, is that persona a good one, or is it annoying?

The assistant can be viewed as a separate entity, one which is primarily conversational and has a "personality." It interprets a user's input and mediates between the user, the native user interface of the device, and a variety of on- and off-board applications. In some cases, the assistant even renders retrieved information in a reformatted fashion within its own UI – and in this sense acts as a lens and filter on the Web.

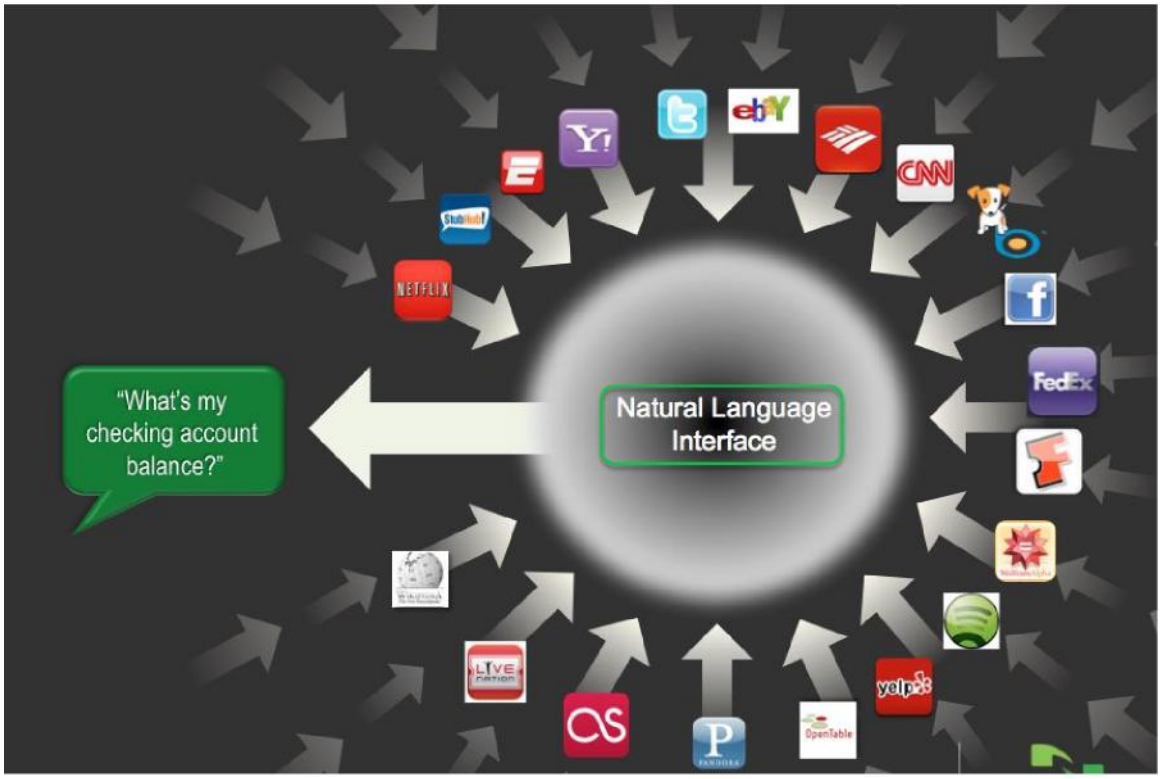
An alternative design, which may be termed "ambient NLU," retains the look-and-feel of native device and application interfaces, but "embeds" context-sensitive NLU. By speaking to this interface, the user can retrieve information as well as open and control familiar-looking applications. The system engages in conversation when needed to complete multi-turn transactions or resolve ambiguity. Rather than being front-and-center, the NLU aims to be unobtrusive, efficient, and open. To the extent possible, it accomplishes tasks based on a single utterance and never limits where the

user can get information. In contrast to the assistant, which aims to help the user address the shortcomings of the existing UI, ambient NLU holds the promise of becoming an inherent part of an improved UI.

In either case, voice and language understanding can now be viewed as a new fundamental building block – a new dimension added to the traditional visual UI, one that permits access and manipulation of assets that may not be visible, located either on the device or in the cloud. Over the next few years, we will see active exploration of the best way to make use of these new dimensions and a rapid revision of the current "shrunken desktop" metaphor as designers experiment with new ways of structuring the experience.

A Network of Services − and Devices

NLU is expanding the boundary of the traditional voice interface by transparently incorporating other services – the apps and content that it invokes – and further supporting the notion that devices are quickly transforming into intelligent systems. As they themselves become voice-enabled, they create a continuity of experience, amplifying the value of NLU UIs as well as of the Web. They allow users to locate and invoke a broad range of services spanning entertainment, travel, shopping, and social media – a step in the direction of the Semantic Web, where agents can discover and interact with services on behalf of the user.



Users increasingly expect to be able to perform similar tasks and access similar information on an increasing number of device types. Another value multiplier is thus an emerging notion of continuity of experience across smartphones, tablets, Ultrabooks, and cars, as well as TVs. While their display form factors and CPU power differ, these

devices all have substantial on-board computation and sophisticated displays, and are all connectable to the Web.

While the manufacturers of these devices vie for the predominance of their particular platforms, users are now dividing their attention among all of these, and increasingly interact with multiple devices every day. This has exacerbated the need for a "portable" experience – a continuity of the functionality, the interaction model, the user's preferences, and interaction history across these devices. Within a few short years, we will see devices use voice biometrics to recognize users, to access a server-based user profile, to adapt it during use, and to update it back on the server after the session ends, making it available for access from the next device with which the user interacts.

Mobile Devices Becoming More Aware

The NLU interface is agnostic with respect to how a user finds it most convenient to convey information – through voice, touch, or gesture. The output of modality-specific meaning extraction all feeds into a single reasoning system, which can also take into account a vast amount of metadata to determine a user's location, rate of travel, and time of day, as well as preferences and usage history, in order to determine the optimal response to the input. These devices are contextually aware and are getting to know our wants, needs, and desires – and more important, our preferences.

In many cases, it can be advantageous to process a number of input streams simultaneously. A video stream from user-facing cameras, which are now a standard feature on most smartphones, can drive a lip-tracking algorithm that can improve the accuracy of voice recognition in noisy environments or help direct the steerable microphone beam toward the user.

Chip-makers increasingly recognize the importance of these novel interface modalities and are developing specialized processor architectures optimized for multiple-input sensors, as well as voice and NLU processing. These include low-power, digital-signal-processing–based "wake-up-word" recognition, which allows users to speak to devices without having to first turn them on, further reducing the number of steps separating user intent and desired outcome. For instance, the Intel-inspired UItrabook is integrating these capabilities, responding to "Hello Dragon" – and from there is ready to listen to a user's command or take dictation.

Other enhancements include built-in noise suppression and echo cancellation (which allows users to speak while the device is playing prompts or music) and the introduction of co-processors to accelerate the core likelihood computation used by voice-recognition engines. Such hardware optimization is essential to fulfilling the promise of voice input and NLU.

Devices are thus becoming more aware of our environments, and by extension, the device interfaces are becoming smarter about who we are, where we are, and what we are doing, and thus what we want them to do for us.

The Car: Just Another Mobile Device?

The automotive market in particular has become a hotbed of interface innovation. Some industry insiders have actually likened the modern car to a mobile device. As Ford Motor Company CEO Allan Mulally put it, "*We're going to be the coolest, most useful app you've ever had, seamlessly keeping you connected."*

According to IMS Research, the world market for connected cars will grow 650%, to reach 40.5 million unit sales in 2017. Additionally, in J. D. Power and Associates' "2012 U.S. Automotive Emerging Technologies Study" on consumer demands for in-car technology, 69% of respondents are reported as wanting natural language voice activation, and 68% want wireless connectivity.

The eyes- and hands-busy usage model makes conversational and seamless access to multiple services a natural fit, with a properly designed NLU user interface providing minimal distraction. The noisy in-car environment, however, presents substantial challenges, and has been the principal driver of the development of sophisticated signal enhancement techniques, including the use of microphone arrays and dynamically steerable beams that zero-in on the desired speaker and block out interference.

In addition, core services need to persist even without network coverage, which has led to the development of distributed systems with voice and NLU technology embedded in the in-car platform as well as located in the cloud, working together to divide up a full range of tasks in a manner that is transparent to the user.

New car models are already hitting the market with message dictation capabilities, such as those in the BMW 7 Series; Audi has also announced plans to incorporate such capabilities. And in 2013 and beyond, these capabilities will rapidly expand to allow people to speak to find local business information and directions, music, and much more, from content in both the car and in the cloud.

Voice in the Livingroom

Voice and NLU are having a profound effect on the digital livingroom, as well. An estimated 1.6 billion TVs will be connected by 2014, according to *Strategy Analytics*. No menu or repository of content is more difficult or more complex to navigate than today's TVs. Traditional television remotes can't keep up, and many find it frustrating having to tap in letters and numbers in search of a program or movie

TV manufacturers and cable operators are currently evaluating ways to integrate voice as part of the user interface. Interactive TVs were among the top highlights at the Consumer Electronics Show 2012. LG and Samsung, among others, are making it easier and faster for people to find and discover new content by simply leaning back and speaking the name of a show, a movie, or even a genre or favorite actor.

The "lean-back-and-speak" experience can be realized by speaking into a microphone on the remote, speaking into an app that turns a smartphone into a remote, or, even better, just sitting back and speaking openly to the television. The latter, also known as "far talking," uses the same beam-forming technologies employed by automotive

manufacturers to filter out interfering noise in the car. As a result, people can talk to their TVs from just about anywhere in the room.

The Transformation of Customer Care

The traditional automated "interactive voice response" industry has not been immune to the rise of the natural language interface. Very much in line with the general "consumerization of the enterprise," users have new expectations of natural interactions with contact centers. By and large, these are still built as telephony voice-only directed dialogs, even though the majority of callers are now connecting to them from multimodal smartphones. Here, too, NLU – for example, in the form of Nuance's virtual assistant Nina – has the potential to cut through deep menu structures and create a more pleasant and engaging experience

Voice and NLU will transform other aspects of this market, as well. The same techniques needed to build semantic search and conversational agents can be applied to high-accuracy retrieval of answers to natural questions from company FAQ databases, well beyond what is achievable with simplistic keyword searches prevalent today. In addition, because it is agnostic to modality, the NLU can be applied to user communication across a variety of channels – not only direct interactions utilizing voice, but also chat and email, thereby allowing companies to understand the full breadth of their customers' communication, and in turn serving them better.

Mobile Voice and NLU in Healthcare

Voice recognition is already being widely used by doctors to capture the complete patient note, often at the point-of-care. Voice input offers a variety of benefits for patients and providers alike, including improved turnaround time of the complete patient record and major cost savings associated with traditional medical transcription.

The industry is also seeing a major increase in the use of mobile technologies as a means to improve care and physician efficiency. Manhattan Research estimates that 62% of doctors now use tablets, with half of those using them in clinical settings. Here, too, voice recognition is playing a key role in enabling usability of mobile devices. Instead of typing and clicking on a small mobile screen, medical professionals can dictate into their mobile device and navigate screens quickly with voice commands rather than keystrokes.

While information entry has become more efficient, government mandates to make the vast amounts of clinical information more broadly available in the form of structured

electronic health records (EHRs) has been enormously taxing for the healthcare industry. Pressures continue to mount as structured data from clinical documentation is now needed to comply with both clinical initiatives and federal billing regulations.

The University of Pittsburgh Medical Center is addressing this challenge by combining voice recognition with clinical language understanding (CLU) – a healthcare-specific form of NLU that utilizes medical-specific ontologies and language-understanding

features – to convert "narrative" medical reports into structured data. This data can then be leveraged to gain insight into inefficiencies and to spot large population health trends. The same NLU framework can tell whether a physician is providing sufficiently complete and consistent information, and if not, prompt for clarification. This improves the quality of the documentation. The improved patient record is then automatically processed further to assign standard billing codes that ensure appropriate reimbursement for the care provided.

We have reached a major turning point in healthcare, as these technologies make critical information more readily available than ever before. It is not an exaggeration to say that a more efficient distribution of key medical information could in many cases have as profound an impact on the quality of patient care as fundamental breakthroughs in medical science.

So... Is It a Revolution, or an Evolution?

The use of mobile voice technologies exploded in 2012, and with that, the consumer expectation of how devices should understand and react to our needs.

The performance, functionality, and versatility of core voice processing and natural language technologies have also been improving at a rapid pace, and there is every reason to expect that to continue.

When this progression in voice and NLU is combined with accelerants in the form of ever-more powerful devices with new sensors, ubiquitous fast networks, vast quantities of new content, services and applications, and users' demand for constant connectedness and communication, we see something akin to a wildfire sweeping through entire industries.

Voice and NLU are breathing life into mobile devices and applications, giving them personalities and the ability to help us in unexpected new ways. We are entering an era of intelligent systems that interact with users in very rich ways, understand what the input means in real-world contexts, and then solve high-value tasks using advanced reasoning.

This truly is a revolution: voice and NLU are now widely accepted as essential elements of the modern user interface, and the way we engage our devices, electronics, apps, systems, and services will never be the same.

About the Author

As Nuance's chief technology officer, Vlad Sejnoha oversees the company's research and focuses on core technology and product strategy with an emphasis on emerging areas, including natural language processing and mobile applications.

Prior to joining Nuance, Vlad was the chief scientist at L&H, and earlier at Kurzweil AI, where he was responsible for creating technology for a number of commercially successful speech recognition products, including large-vocabulary continuous-speech dictation systems.

Vlad has over 20 years' experience in the field of speech recognition and holds 13 US patents.