

Revolutionizing Voice UI for Mobile

Vlingo Unconstrained Speech Recognition White Paper



Vlingo User Interfaces

Vlingo is focused on solving the overall usability challenge on mobile phones. As devices, networks, and applications are improving, mobile phones are becoming people's primary mobile communication, entertainment, and information device. Wireless operators, device makers, and application providers are all finding that the limitations of a small display and twelve buttons on a keypad is becoming the primary constraint on what applications can be successfully deployed. We believe that a more capable user interface is the key to unlocking data services for a broad segment of the population.

Speech recognition has been used to varying degrees of success in a variety of mobile applications such as voice dialing. The adoption of speech as an interface on mobile devices has been limited however by the constrained nature of these speech-driven applications. Applications such as voice dialing, directory assistance, and content search have been characterized by:

- **Constrained Grammars:** In order to achieve reasonable accuracy, most speech recognition interfaces operate with a constrained grammar at each step in the interaction. Some of these have been simply a list of allowable words or phrases (the names in the user's address book for example). Even in the case of applications which advertise a "natural language interface", the speech recognizer is still constrained to a grammar of allowable patterns and words ("call <name> on <his/her> <mobile/office/home> number").
- **Scripted Interaction:** There have been systems which go beyond simple functionality to perform tasks such as 411-style directory assistance. By using network side processing, these systems have allowed the use of much larger grammars. But, these grammars are still constrained and cause the systems to lead the user through a number of steps to accomplish the targeted task. For example, in the case of directory assistance applications, the system first gets the user to speak the city/state and then asks the user to speak the name of the business (this is so the system can load a grammar of business names for the specified city/state).

The need to develop these constrained grammars and scripted interactions imposes a significant burden on the application developers since these systems need to be carefully designed and tuned to have successful interactions given these constraints. These tasks require both intensive manual effort as well as significant expertise, resulting in the requirement of specialized speech-recognition experts to develop usable speech applications. Costs for these experts start around \$100k for a small low-end application, and quickly escalate into the millions of dollars.

More importantly, in order to successfully use the application, the end users need to learn about the constraints (since the applications fail if users speak things outside of the constrained grammars) and how to navigate the scripted interaction. Since these are different for each grammar-based application, users need to learn this for each of the voice enabled applications that they use. This is usually a trial and error process. Perhaps more accurately, it's an error and retry process. Clearly this does not scale across all of the applications people will want to use on their mobile devices.

The type of error caused by users speaking unexpected things is commonly known as "out of grammar", and is the most frequent cause of all misrecognitions. Major speech vendors commonly report accuracy figures only for in-grammar sentences, ignoring the out-of-grammar problem. When the definition of accuracy is expanded from *in-grammar accuracy* to *user-perceived accuracy*, it becomes easier to see why traditional speech implementations have not always been warmly embraced by users. With its already high accuracy on expected utterances and properly handling of unexpected utterances, Vlingo has fared quite well in competitive

technology bakeoffs. In fact, the harder the task, the more the benefits of our HLM-based approach become apparent. We've found the best way to understand vlingo's accuracy is to try it for yourself.

Vlingo Approach

In contrast to these existing systems, vlingo has gotten rid of the application-specific grammar constraints. This has in turn removed the need for the scripted interactions. Taken together, these simplifications have allowed vlingo to pioneer dramatic cost savings that fundamentally change the economics of top-tier speech systems. Now even the smallest and newest developer can add afford the time and cost to speech-enable their mobile applications.

Instead, vlingo has a very simple approach from the user's point of view – they should be able to type or speak anything they want into any vlingo-enabled text box, at any time. An example of this can be seen in figure 1.

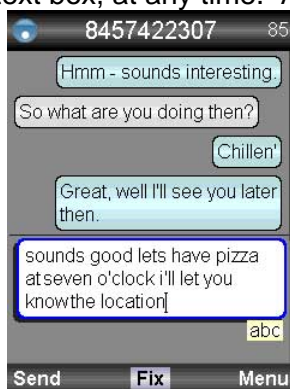


Figure 1: Example of vlingo-enabled application. This Text Messaging application allows the user to type or speak any text into any text entry box in the application.

While this is a simple concept, it is very challenging to achieve (removing the grammar constraints and allowing users to speak anything they want puts a huge burden on the underlying speech recognition technology).

We are able to successfully achieve this using:

- **Hierarchical Language Model Based Speech Recognition:** We have replaced the constrained grammars with very large vocabulary (millions of words) Hierarchical Language Models (HLMs). These HLMs are based on well-defined statistical models to predict what users are likely to say given the words they have spoken so far (“let’s meet at ___” is likely to be followed by something like “1 pm” or the name of a place). While there are no hard constraints, the models are able to take into account what this and other users have spoken in the particular text box in the particular application, and therefore improve with usage. Unlike previous generations of statistical language models, the new HLM technology being developed by vlingo scales to tasks requiring the modeling of millions of possible words (such as open web search, directory assistance, navigation, or other tasks where users are likely to use any of a very large number of words).

- Adaptation:** In order to achieve high accuracy, vlingo makes use of significant amounts of automatic adaptation. In addition to adapting the HLMs, the system adapts to many user and application attributes such as learning the speech patterns of individuals and groups of users, learning new words, learning which words are more likely to be spoken into a particular application or by a particular user, learning pronunciations of words based on usage, and learning peoples' accents. The adaptation process can be seen in figure 2.

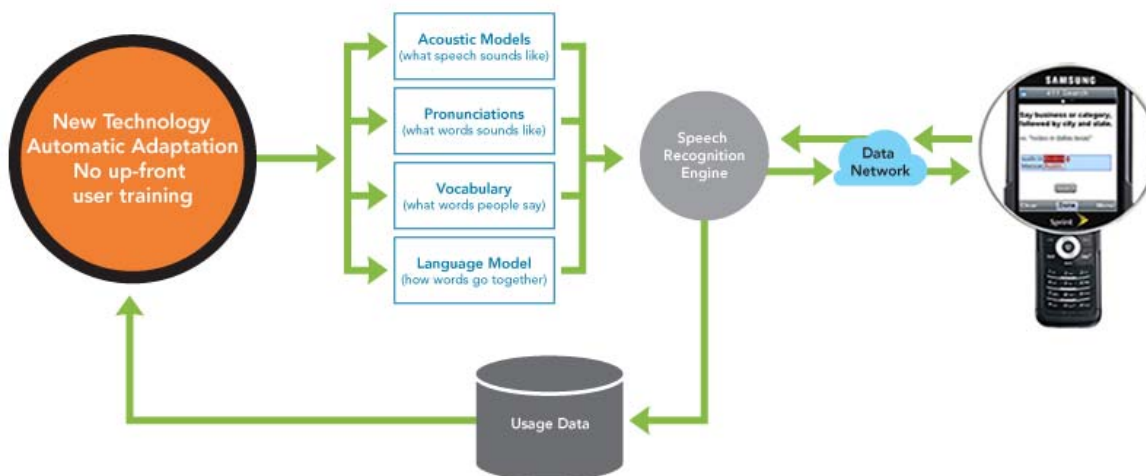


Figure 2: Vlingo Adaptation Architecture. The core Speech Recognition engine is driven by a number of different models, each of which is adapted to improve its performance based on usage data.

- Server-side Processing:** The vlingo deployment architecture uses a small amount of software (about 50KB – 90KB, depending on platform) on the mobile device for handling audio capture and the user interface. This client software communicates over the mobile data network to a set of servers which run the bulk of the vlingo processing. While this does make our solution dependent on the data network, it enables the use of the large amounts of CPU and memory resources needed for unconstrained speech recognition, and more importantly allows the adaptation described above to make use of usage data across all users.
- Correction Interface:** While the techniques described above result in high accuracy speech recognition across users, there are still errors made by the speech recognizer. In addition, there will be situations where the user will prefer to enter text using the keypad on the phone (where they need privacy, in high noise environments, or when the speech system is unavailable due to lack of network coverage). Therefore, we have designed the user interface to allow the user to freely mix keypad entry and speech entry (at any time the user can either type on the keypad or push the “talk” button to speak), and to allow the user to correct the words coming back from the speech recognizer. Users can navigate through alternate choices from the speech recognizer (using the navigation buttons), can delete words or characters, can type or speak over any selected word, and can type or speak to insert or append new text wherever the cursor is positioned. We think this correction interface is the key to allowing users to feel confident that they can indeed efficiently enter any arbitrary text through the combination of speech and keypad entry.

Response Time

Vlingo's servers are designed to respond to user requests in under half a second (0.5s). However, this does not tell the whole story. The more important metric is latency as perceived by the user. Here, there are several key variables, including:

- Whether the mobile device can stream audio in real time, or can only begin to send audio after the user finishes speaking.
- Network speed.
- Length of the spoken utterance.
- Processing speed of the mobile device itself

The effect of these variables is substantial. At the fast end of the spectrum, typical latency of a benchmark application running on a smart phone that supports streaming audio over a 3G network is 1-2 seconds. The real-time audio streaming capability more than makes up for the modest network speed, regardless of the length of the utterance. In the midrange, a mid-level 3G feature phone which does not support streaming will see a latency of roughly 2-3 seconds for the same application. At the slow end, a popular low-cost 2.5G feature phone typically turns in response times in the 4-5 second range.

Data Requirements

Vlingo uses the standard audio encoding available on each mobile device. The encoded audio is sent from the client to our servers, where all of the recognition processing takes place. The two standard encoding schemes, AMR and QCELP, each have bit rates around 11-12 kbps. The duration of a typical utterance is 2 to 5 seconds, depending on the application type. So, the total data transported will range anywhere from 22 to 60 kilobits, or approximately 3 to 8 kilobytes per utterance.

Recognition results are sent back from the server to the client device. Since the results are simple ASCII text, their size is negligible compared to the encoded audio.

Applications

Unlike the grammar-constrained systems, the vlingo approach allows truly multi-modal (speech plus keypad) interactions with any mobile application which requires text entry.

The sorts of applications which are now possible include:

- **Messaging Applications (SMS, Email, Social Networking, IM, Blogs/Chat):** These applications are clearly not possible with grammar-based speech recognition due to the very wide variety of input needed, and are very well suited to the vlingo interface. Due to the amount of text entry needed, we think that these are the applications which are mostly likely to benefit from the ability of users to speak their input.
- **Search (content search, local search, open web search):** While it may be possible to handle static content search with constrained grammars, as operators and application providers want to extend beyond constrained content search, we think it is important to have a single consistent interface which operates across domains. The vlingo approach is able to scale from the constrained case to cross-domain open search with no changes to the architecture or user interface. Furthermore, many search

tasks such as music or news searches require rapidly-changing vocabulary domains that are difficult to keep current with manual grammar edits, but easy to handle with an automated adaptation approach.

- **Navigation/Location Based Services:** GPS-enabled navigation systems on mobile phones along with associated location based services provide very compelling functionality to end users. However, the experience rapidly degrades when the user needs to enter an address or search for a business as a destination. This can be a tedious experience even when users give their full attention to the interaction, and can be very dangerous when people attempt this while driving (which they will do despite warnings provided by the interface). In fact, predictive text entry systems such as T9 will not have many business names in their dictionaries, further degrading the experience. Simply being able to speak something such as “17 Dunster Street Cambridge Massachusetts” into the destination field is a much easier, faster, and safer interaction than entering this through the keypad. Further, the emergence of combined navigation/local search applications can only work with an “open” voice interface such as that offered by vlingo, since users are likely to want to enter either a street address or a business name. With hundreds of new businesses opening every day, a grammar-based speech recognition system simply would not work.

In addition to being able to enable all of the above applications, we believe a key feature of the vlingo interface is that we can enable all of the above in a *consistent* interface. So, once a user becomes accustomed to the interface in one application (and the system has had a chance to adapt to the user’s speech patterns), they can use what they have learned to successfully use another vlingo-enabled application.

Products

Vlingo is first and foremost a speech recognition service. In addition to our speech recognition service which is the core of everything we do, we offer three types of products to the market:

- A **Software Developers Kit (SDK)** which is used to vlingo-enable any text box in any mobile application. The key component of the SDK is a device-based software library with an API which replaces the device’s native text box with a vlingo text box that has all the existing native functionality, but is also vlingo-enabled. The SDK is offered to application developers for integration with their individual applications, and to wireless carriers and device manufacturers to vlingo-enable any application on a phone or line of phones.
- **Vlingo voice**, a unique application that addresses not only usability but also the problem of discoverability of mobile content, by allowing users to speak virtually any request and interact with any application directly from their phone’s idle screen without having to search through the phone’s menu structure or wait for the application to start up. This is made possible by vlingo’s HLMS, which have the unique ability to combine multiple functions in a single text box without significantly degrading accuracy. For example, using vlingo voice, a person can press a single button and speak any of the following requests:
 - “Call Jen Jones on her cell phone”
 - “Send email to Pete, subject ‘Running late’, message ‘Pete, I’ll call you when I get to the office’”
 - “Web search, ‘Italian restaurants in Boston’”
 - “Check the weather in Seattle”
 - “Open calendar”
 - “Send a text message to Mom, ‘Happy Mother’s Day’”
 - And many other popular functions that a user can perform from their mobile device.

Additionally, rather than forcing users to follow a pre-defined command grammar, or structure, vlingo voice adapts to the way that users naturally speak their requests.

- **Mobile applications.** Vlingo has developed several voice-enabled applications, including:
 - FIND, a local business search and mapping application.
 - Voice2Txt, a text messaging application with threaded conversations akin to instant messaging.
 - A web search application that allows users to search the web by voice.
 - A video viewer which searches for and plays YouTube videos.
 - Several other demonstration applications.

Summary

Vlingo is taking a very different approach to speech interfaces than what has been deployed to date. We think this new approach has compelling benefits to the end users and to the organizations which are developing and deploying applications.

The vlingo approach has several unique benefits which are self-reinforcing given the nature of our architecture:

- “Open” voice recognition architecture without application-specific grammar constraints: Other speech recognition approaches will not allow users to speak naturally, but can only recognize certain predefined words that are “allowed” by the grammar. Vlingo’s open approach also makes it possible to speech-enable any input into mobile applications.
- Adaptation from server-side processing: With server-side processing, applications are updated in real time and are self-healing. For example, hundreds of new businesses open every day and a phone-based, grammar constrained speech recognition system simply would not recognize these business names. In the vlingo approach, if a user speaks a new business name, that name is learned by the system, and the language model is updated for all users of the service.
- Multi-modal input with correction interface: Users will continue to demand multi-modal input. Text input is required in certain situations; however, a voice user interface unlocks the power of data services on today’s feature phones that are so user-interface restricted. The ability to mix text with typing allows users to correct any errors that might emerge when they are speaking with an accent or speaking a new word. Vlingo’s architecture captures this correction, and the system learns so that other users with a similar accent or also using the new word benefit.
- Substantially reduced deployment cost: As a result of eliminating application-specific grammars and their associated scripted user interfaces, applications can be speech-enabled with only an API integration to vlingo products rather than costly speech-specific tuning work requiring scarce speech-recognition expertise.

Because we believe this interface is broadly applicable, we are working with carriers and application partners across multiple application domains to incorporate the vlingo interface and deploy to broad sets of end users.

In addition to our core recognition service, vlingo products include an SDK for integrating vlingo into third-party mobile data applications, vlingo voice which addresses both usability and discoverability by allowing users to speak any request directly from their idle screen, and a number of single-purpose applications such as FIND and Voice2Txt.