Chapter 2
“Striking a Healthy Balance”: Speech Technology in the Mobile Ecosystem

Scott Taylor

Abstract Mobile speech technology has experienced an explosion of adoption across a variety of markets – from handsets and automobiles to a variety of consumer electronic devices and even the mobile enterprise. However, we are just scratching the service on the benefits that speech can provide not only consumers, but also carriers and manufacturers. This chapter takes a closer look at the advent of speech technology in the mobile ecosystem – where it is been, where we are today, and where we are headed – keeping in mind the delicate balancing of a variety of multimodal capabilities so as to optimally fit the user’s needs at any given time. There is no doubt that speech technologies will continue to evolve and provide a richer user experience, enabling consumers to leverage the input and output methods that are best suited for them moment to moment. However, the key to success of these technologies will be thoughtful integration of these core technologies into mobile device platforms and operating systems, to enable creative and consistent use of these technologies within mobile applications. For this reason, we approach speech capabilities on mobile devices not as a single entity but rather as part of an entire mobile ecosystem that must strive to maintain homeostasis.

Keywords Mobile ecosystem • Multimodal navigation • Multimodal service calls • User experience • Speech technologies • Integration into mobile device platforms and operating systems • User interface challenges to designers • Hands-free • Enterprise applications and customer service

S. Taylor
Vice President, Mobile Marketing and Solutions, Nuance Communications, Inc.,
1 Wayside Road, Burlington, MA 01803, USA
e-mail: Scott.Taylor@nuance.com
2.1 Introduction

The availability of computing power and network connectivity in automobiles, mobile phones, and other mobile devices has led to an explosion of available applications and services for consumers. Maps and navigation, the advent of social networking sites like Twitter and Facebook, email, web search, games, and music and video content have become commonplace on mobile devices, and are now emerging as services available in cars and in other electronic devices.

But as these new services and applications become more popular, they pose many user interface challenges to designers. For instance, devices are limited in computing power, display size, and the keyboard is small and difficult for many people to use. Also, the convenience of mobility creates situations where the users are not always able to keep their eyes and hands on the device—walking, engaging in conversation, working out at the health club, and the obvious—driving a car. With these challenges in mind, device manufacturers have invested heavily in technologies that ultimately improve the user interface experience, such as predictive text, touchscreens, and speech technology.

Speech technologies, including both speech recognition and text-to-speech, have been popular for use in mobile applications for decades. However, until recently, that popularity was limited to niche applications, such as voice-dialing or assistive applications, to help the disabled use mobile technology. In the last few years, there has been a rapid expansion in the breadth of mobile applications, leading to an increased demand for speech technology.

Historically, speech technologies have been preloaded on devices at the factory—on mobile phones, in automotive in-car platforms, or in gaming devices. It is available in the device right off the shelf. However, as third generation (3G) and fourth generation (4G) wireless data networks have become prevalent and more robust, application providers are now using the additional computing power that is available to provide more advanced speech capabilities to mobile devices and downloadable applications.

Today, mobile speech applications tend to be focused on core services such as navigation, dialing, messaging, and search. In the future, we will see speech used in a wide variety of mobile applications, including entertainment, social networking, enterprise workforce, mobile banking and payment, customer service, and other areas. Speech technology will also become available in a wide array of devices.

2.2 The First Mobile Voice Applications

2.2.1 Voice Dialing and Voice Commands on Phones

One of the first mobile voice applications to emerge in the 1990s was voice dialing, which allowed users to press a button and speak a number or name to call so that the user could place phone calls without looking at the keypad and trying to find numbers.
Initial voice-dialing technology used speaker-dependent technology, or “voice tags.” With speaker-dependent technology, the user of a device needed to go through an enrollment process, where they would speak recordings of the digits and names that would be used for dialing. Each digit or name typically had to be recorded one to three times, and the voice-dialing application would only work for the user who enrolled.

One advantage of speaker-dependent dialing is that it was language independent. A device equipped with “voice tag” capabilities could be used by a speaker in any language. However, the voice-dialing applications could not automatically learn new names as they were added to the contact list, and the enrollment process was frustrating for many users. Unfortunately, many users formed early and negative impressions of mobile speech recognition capabilities from these early speaker-dependent systems. Today, as the technology evolves, it continues to be a challenge for the industry to overcome those negative first impressions.

Computing power and memory footprint continued to increase on mobile devices. Device manufacturers soon added more sophisticated phonetic speech recognition capabilities to the device. Phonetic speech recognition used acoustic speech recognition models trained on a wide variety of speaker voices and styles, and recognized phonemes rather than word templates, and had the following advantages:

- No user enrollment or training was required.
- New words and contact names could be added dynamically. If a new contact name is added to the contact list, then it could be recognized by the voice dialer using standard phonetic pronunciation rules.
- The application could recognize flexible manners of speaking. For example, a user could say “Call John on his mobile,” or “Dial John Smith on his cell phone.” If the application was correctly programmed, it could handle a great deal of flexibility.

Some voice command applications could also be programmed to recognize a long list of commands, beyond just dialing. In fact, some phones today can recognize 50–100 voice commands to control the device. Popular hands-free commands include:

- “turn Bluetooth on/off”
- “send text to <contact-name>”
- “check battery”
- “check signal”
- “go to camera”
- and more

Unlike voice tags, phonetic speech recognition on the phone was not speaker dependent, but rather language dependent, meaning that the software works out of the box for any user, but only recognizes specific languages and dialects. With that in mind, it became very important for on-device technology to support many languages given today’s global landscape. And while this language-dependent technology can support a variety of commands and speaking styles, it nevertheless requires the user to use gate commands like “dial Jim” or “go to voicemail.” In this instance, users must have a sense of which commands are supported on the device – potentially creating an additional learning curve for some users.
2.2.2 The Advent of the Hands-free Experience on the Phone

Voice dialing and other voice commands were expected to work well in situations, where the user’s hands and eyes were not completely free, and so it was important that these applications provided a minimal attention interface.

Implementers of speech recognition systems on a device needed to consider the amount of button pressing and holding required to use speech recognition. The simplest and safest interfaces required only a simple button push, as described in the following sequence:

- User pushes a button and quickly releases it to activate voice commands
- The voice command application prompts the user via an audio cue to begin speaking
- The user says, “Call Jim on his mobile”
- The voice command system automatically detects when the user has finished speaking and begins the dialing process
- If any disambiguation is required (for instance, if there are multiple entries for “Jim”), the voice command system resumes listening without requiring another button push from the user

To detect when the user starts and stops speaking, the speech recognition technology had to perform a technique called “endpointing.” Endpointing had to be carefully implemented, in order to avoid interrupting the user when they pause briefly while speaking. Careful testing of the audio interface on the device was required. Not all speech recognition systems supported endpointing because of the complexity of the algorithms and the need for close integration to the device.

It was also important for these speech dialers to provide audio cues to the user for when they were not looking at the device. Audio prompts and high quality text-to-speech have been incorporated into some applications to provide audio confirmation of the name/number being dialed, and to disambiguate if there are multiple matches. For example:

User: “Call Jim on his mobile phone”  
System: “Multiple matches found…Jim Ardman…Jim Smith…Jim Workman”  
User: “Jim Smith”  
System: “Calling Jim Smith’s Mobile Phone”

Text-to-speech must be used in this example to playback names from the contact list. If high quality text-to-speech is embedded on the device, then it can be used to enhance the minimal attention interface by performing services such as:

- announcement of incoming caller ID number or name
- announcement and reading of incoming text messages
- announcement and reading of incoming emails
- reading menus aloud

For the last several years, device manufacturers have been deploying the applications with phonetic speech recognition and high quality text-to-speech. One
example is the Nuance’s Vsuite product, which can support dozens of languages and contact lists with thousands of names. These applications perform best when integrated as a fully integrated capability of the device, in order to provide the best possible user experience.

2.2.3 Drivers Begin Talking to their Cars

Several years ago, auto manufacturers began putting computing platforms into cars to add new features to the car, including voice commands. Typical voice commands have included Bluetooth-enabled voice dialing, and voice control of in-car functions, such as turning the radio on/off, changing stations, changing CDs, or modifying the heat/air conditioning temperature settings. Text-to-speech technology has also been used to provide turn-by-turn driving directions for in-car navigation systems, as well as after-market navigation systems that can be installed by the car owner – like those offered by TomTom and Garmin. In recent years, navigation applications have even incorporated more sophisticated speech capabilities that allow users to enter destinations (addresses and points of interest) just by using their voice, with full step-by-step confirmation with the use of text-to-speech technology.

The automotive environment presents one of the most challenging environments for speech recognition. It is essential to minimize the visual and manual engagement required by the driver: there can be many passengers speaking in the car while commands are given, or there can be music playing, or there can even be simpler elements of background noise coming from outside, such as wind and other factors.

For these reasons, automotive manufacturers have invested in the optimization of speech applications for a specific car environment. They have incorporated high-quality built-in microphones and noise reduction technology. Applications were trained on audio data using the specific acoustic environment of the car.

2.2.4 Assistive Applications on Mobile Devices

Speech technologies have been used on mobile devices to enable and enhance service for blind and visually impaired users, as well as those in the disabled community. Common applications included:

- voice dialing with audio confirmation
- screen reading
- caller ID announcements
- reading incoming text messages and email

Assistive applications needed to consider the needs of the community of users carefully. For example, Nuance Communications TALKS screen reader for mobile devices included features for adjusting the volume and speaking rate of text-to-speech, and also included integration with external Braille input/output devices.
2.3 Speech Technology and the Data Network

As described in the previous section, speech recognition and speech synthesis can be performed on mobile devices with great success, and the technology has continued to get better from year to year. However, speech technology is hungry for CPU and memory cycles. The emergence of higher powered devices provides more processing power for on-device speech; however, these devices also come equipped with many new services such as web browsing, navigation and maps, and media players that consume resources – but do create a need for much more advanced speech recognition than traditional voice dialing or commands.

Fortunately, the availability and reliability of wireless data networks is rapidly increasing, and many of these higher-end devices are equipped with unlimited data plans. This creates a great opportunity for speech, allowing speech-based applications to take advantage of the data network to perform advanced speech processing on network-based servers rather than on the device itself. With network-based speech recognition, the audio is collected on the device by the application, and transmitted across the data network to specialized servers that perform transcription of audio to text and then sends the text back to the device. With network-based text-to-speech, the text is sent to servers and converted to audio which is streamed back to the device.

Network-based speech technology has several key advantages, namely,

- speech technology can take advantage of unlimited processing power in the cloud
- with this computing power, tasks such as speech-to-text transcription can be done very accurately
- some tasks, such as web search and navigation, can take advantage of data on the network to improve accuracy (web listings, address listings, movie names, etc.)
- the technology can be easily refreshed on the server side so that it stays up to date, whereas “factory installed” technology is usually not updated in the field
- speech that is processed in the network can help researchers improve accuracy of the core technology

There are, however, some limitations:

- Highly-used networks can introduce latency. If the network is fast and not congested, then results may typically be returned in a few seconds. However, if the network is slow or experiencing a high volume of usage, results may take much longer
- the data network is not yet highly available in all areas
- if the speech application itself is not factory installed, it may be more difficult to capture audio effectively and to integrate seamlessly with applications on the device
- some applications, such as voice dialing or music playing, if implemented on the network, would require that users upload personal data to the network

In the next 5 years, we can expect to see hybrid solutions that leverage both on-device and off-device speech resources performing most effectively in the mobile environment. For example, a device may leverage local resources for voice dialing and local music playing, and go to the network only when it needed resources for speech-to-text dictation or voice search.
2.4 Emerging Mobile Voice Applications

In the last few years, a variety of new applications for mobile devices have emerged that leverage network-based speech technology. In some cases, these applications have been made available for download to high-end smart phones such as iPhone, Blackberry, Android, Symbian, or Windows Mobile devices. In other cases, they are preloaded on mobile devices or into automotive platforms.

2.4.1 Voice Navigation and Mapping

Application providers that make navigation and mapping technologies have been among the first to incorporate advanced speech technologies into their applications. Speech technology is used to make input/output easier when on the go, or when using a small footprint keyboard or touchscreen keypad.

These applications can be enhanced by:

- entry of destination address by voice
- entry of landmark or point of interest by voice
- lookup business names or other content criteria (e.g., “Dave Matthews concert”)
- playback of specific turn-by-turn directions using text-to-speech

Implementing speech enabled navigation can be a complex task, especially for multilingual systems. Generic speech recognition technology alone is not enough. The technology must be trained on the “long tail” of addresses and locations for which people will need directions. Also, it is essential that the application support natural language interfaces, as users will have low tolerance for following several steps to input city, state, and to speak the names of businesses or destinations in a highly constrained fashion.

2.4.2 Message and Document Dictation

The emergence of text-messaging and email as popular mobile applications has been rapid, driven in part by the availability of full QWERTY keyboards on mobile devices. However, the keyboards are small and difficult to use for many users, touchscreens are difficult to use for typing, and it is impossible and unsafe in on-the-go situations.

For years, the dictation of text has been a successful application in the desktop and laptop world, with software like Dragon Naturally Speaking that is trusted and used by millions. Network-based computing power now makes it possible to perform speech-to-text dictation from mobile devices. Nuance has recently released a version of Dragon Dictation for the iPhone that provides a simple user interface for dictating text for email, text messages, social networking applications, and any application that requires text entry.
Dictation technology will work best when integrated into the applications that use dictation, such as email and messaging clients. On some mobile operating systems, such as Symbian and Android, it is possible to include speech as a universal input method that is active in any application that allows text entry. On feature phones and other operating systems, it may only be possible to include speech dictation by modifying the applications that need to use dictation to interact directly with the recognizer.

There are several important ingredients for success of speech dictation in mobile applications:

– the speech-to-text technology must be mature and robust for the language which is being dictated...it can take years of real-world use from a variety of human voices to make this technology robust
– the user interface must be clear about when and how to activate speech recognition
– ideally, the speech recognition technology can learn from the user’s voice, common names they use, common terms used in their email and messages...this can require the user to give permission to upload some of this data to the network
– the user must have some way to correct mistakes; ideally, this will be a “smart” correction interface that gives the user alternate word/phrase choices so they do not need to retype

### 2.4.3 Voice Search

Similar to voice dictation, voice search allows the user to perform search queries using their voice. These queries could be:

– general search queries fed into a search engine such as Google, Bing, or Yahoo
– domain specific queries, such as searches for music or video content, product catalogs, medical conditions and drugs, etc.

For voice search to work well, the speech technology must be trained on common terminology that is used in search queries. A general voice search engine should know about celebrity names, top news topics, politicians, etc. A medical voice search engine should be trained on medical terminology and drug names.

Voice search has been built into many popular search engines. However, it may become more interesting as applications emerge that can determine the type of search and the user intent, and launch searches into appropriate content sources.

### 2.4.4 Speech Applications for Consumer Devices

Speech technologies have been deployed on a variety of mobile devices other than mobile phones and automobiles. Examples include:

– voice commands for portable gaming devices such as the Nintendo DS
– text-to-speech for reading content on mobile content readers such as Amazon’s Kindle
voice recognition and text-to-speech on MP3 music players to play artists, song titles, and playlists

2.5 Speech and the Future of Mobile Applications

2.5.1 Enterprise Applications and Customer Service

Enterprises, such as banks, mobile operators, and retail companies, have begun to invest in mobile applications. The rapid adoption of smart phones, such as iPhone, Blackberry, and Android-based phones, has provided a set of platforms for the development of downloadable applications that can reach a broad segment of the customer base.

Speech recognition provides many benefits to customer service applications today in over-the-phone voice response applications. These benefits can be extended to mobile customer service applications as well so that callers can speak to mobile applications in order to get product information, account information, or technical support. Speech can remove usability constraints from the mobile interface and allow enterprises to build more complex applications that provide better self-service capabilities.

Potential examples of speech usage would be:

- Using an open-ended “How can I help you?” text box at the beginning of the application that would enable the user to type or speak their question and then launch an appropriate mobile applet (a small application that performs limited tasks) that would provide service…instead of forcing the user to navigate through visual menus.
- Adding a product search box to a mobile application, and so the user could say the name or description of the product for which they need service.
- Speaking locations for store/branch locators.
- Speaking lengthy account numbers or product codes for service activation
- Dictating text into forms for applications (e.g., a mobile loan refinancing application).

Companies may find valuable use for mobile workforce applications, such as:

- Dictating notes into CRM applications
- Dictating notes into work order management
- Dictating into mobile order processing applications

2.5.2 Natural Voice Control

Now that it is possible to accurately convert speech-to-text using the computing power available via the data network, it is possible to take the next steps in voice control of devices and applications. Today’s voice command systems present a limited set of choices, and users must have some idea of the syntax used for commands.
As the number of applications and services available on mobile devices expands, it will be necessary to provide a more natural spoken interface for users, and to provide an intelligent interpretation of the user’s intent. For example:

<table>
<thead>
<tr>
<th>User’s request</th>
<th>Appropriate action</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Send text to John Smith…I’ll meet you at Redbone’s at 6pm”</td>
<td>Launch the text messaging client, address the message to John Smith from the contact list, and feed the text into the message client</td>
</tr>
<tr>
<td>“Find the nearest Mexican restaurant”</td>
<td>GPS locate the phone, launch the default maps/navigation software, and search for Mexican Restaurants</td>
</tr>
<tr>
<td>“Call John on his cellphone”</td>
<td>Determine if John is the only “John” in the contact list…if so, then place the call…otherwise prompt for more info</td>
</tr>
<tr>
<td>“Turn my Bluetooth on”</td>
<td>Activate Bluetooth</td>
</tr>
<tr>
<td>“How tall is the Eiffel tower?”</td>
<td>Launch a search application and feed it the text</td>
</tr>
</tbody>
</table>

Translating the spoken words to text is the easy part; determining the actual intent for diverse populations of users, in a variety of languages, is the challenging part. Supporting this type of voice control system for a wide variety of global languages and cultures is also difficult. And finally, integrating voice control into a variety of applications on a wide variety of devices and platforms could be very difficult. However, the technical capabilities exist today, and so certainly mobile devices will evolve in this direction.

### 2.5.3 The Future of Multimodality

Predictive text, speech recognition, and text-to-speech software are already prevalent on many devices in the market. Other technologies are also emerging to make it easier to input or read text on a variety of devices, including:

- Continuous touch technology, such as Shapewriter, which allow users to slide their finger continuously around a touchscreen keyboard to type.
- Handwriting recognition technology, such as Nuance’s T9Write, which recognize characters, entered on a touchscreen.
- Font rendering technology, such as Nuance’s T9Output, which provide capabilities for more dynamic and flexible presentation of text fonts on mobile devices.
- Haptic feedback technology which provides vibration or other cues to the user.

Today, users typically must choose a particular mode of input or output. Traditionally, different input/output technologies have not always interacted seamlessly, though that phenomenon is starting to change, as some devices have begun to combine speech and text input in interesting ways. For instance, Samsung devices like the Instinct and the Memoir allow users to pull up the text input screen...
with their voice, and automatically bring them into a touchscreen QWERTY text input field that features predictive text….however, users still find themselves either in speaking mode or typing mode, but not both at the same time.

There are situations where voice input is not appropriate or not feasible: in a meeting, or at a loud concert, for example. Similarly, there are situations where text input is not feasible or safe: driving a car, walking the dog, carrying packages. It will become increasingly important for input/output technologies to interact seamlessly based on user choice and preference.

For example, consider the following potential multimodal interactions, which could be implemented with technologies available today.

2.5.4 Multimodal Navigation

- The user presses a button and speaks a query: “Find the nearest coffee shop.”
- The application GPS locates the phone, and then launches a map application which presents a map of nearby coffee shops.
- The user uses his finger to draw a circle around the desired coffee shop…the mapping application zooms in on the desired area.
- The user presses the speech button and says, “Send this map to Mike Smith.”
- The email application launches, with a link to the map attached. At this moment, several people walk into the room. The user wants to communicate a private message, and so he uses predictive text technology on the touchscreen to type a message: “I will meet you at this coffee shop at 4:30 to finalize the sales presentation for Acme Corporation. I think we should lower our bid to $400,000…give it some thought.” He then hits the send key.

2.5.5 Multimodal Service Calls

- The user gets a text message from his airline that indicates his flight has been canceled, with a link to a number to call to rebook his flight.
- The user clicks the link and places a phone call and is connected to a service agent, validating the call is from his mobile device.
- The service agent uses a data application to push a set of alternate flight options down to the user’s phone. An application framework on the phone launches while the user is still on the call with the service agent.
- The user can use the touchscreen to scroll through options and look at layover times, seat availability, and arrival times.
- When the user determines the desired flight, he selects the flight.
- The service agent completes the change, and then pushes a boarding pass to the user’s mobile device which can be scanned by the user at the airport.
2.6 Looking Forward

There is no doubt that speech technologies will continuously evolve and provide a richer user experience, enabling consumers to leverage the input and output methods that are best suited for them moment to moment. The key to success of these technologies will be thoughtful integration of these core technologies into mobile device platforms and operating systems, to enable creative and consistent use of these technologies within mobile applications. Continued emphasis on the user experience will also be key, to ensure that users understand where and how to speak to mobile devices in a manner that is successful.
Advances in Speech Recognition
Mobile Environments, Call Centers and Clinics
Neustein, A. (Ed.)
2010, XXVI, 369 p., Hardcover
ISBN: 978-1-4419-5950-8