

Speech recognition technology for mobile phones

Stefan Dobler

Following the introduction of mobile phones using voice commands, speech recognition is becoming standard on mobile handsets. Features such as name dialing make phones more user-friendly and can improve the safety of operating a handset in an automobile.

The author briefly reviews the types of speech-recognition systems, illustrates a few typical problems of the mobile environment, discusses the efficient use of memory in mobile devices, and provides details on the technology deployed in Ericsson's speech-recognition system for mobile devices. Finally, potential enhancements aimed at speech-controlled mobile phones of the future are mentioned.

Introduction

Last year, Ericsson was one of the first mobile phone manufacturers to add an important technology to mobile phones. The T18, launched in spring 1999, was the first commercially available Ericsson GSM phone that could be operated by voice commands using automatic speech recognition, in addition to commands input via the keypad. Other members of the family of telephones using Ericsson's first generation of speech control algorithms are the T28, R320, and A2618 (Figure 1).

These phones use speech recognition for the new *name dialing* feature. Thanks to the efficient use of memory, it is currently possible to train and store voice tags for up to 10 entries in the phone book of any of these phones. Each voice tag is trained with a single utterance by the user and assigned to a single phone book entry. When the user wants to place a call, he pushes a button and speaks a person's name. The phone answers

with the recognized voice tag as acoustic feedback, and then automatically sets up the call.

All Ericsson phones with speech recognition capabilities also feature *call answering*, which allows the user to accept or reject incoming calls using voice commands. This has obvious advantages when the phone is used with hands-free equipment.

Compared to dictation products commercially available for desktop PCs, the application described here seems elementary. However, mobile phones are used every day, in a variety of locations, with every kind of background noise imaginable. Hence, the key issue for speech recognition in mobile devices is not the size of the vocabulary but the robustness of the recognition system.

On one hand, the phone must recognize speech correctly, say, in a quiet office setting, at an airport with conversations going on in the background, or in a car traveling at 150 km/h. On the other hand, care must be taken so that no incidental noise such as a closing door or laughter is mistaken for a valid name, which would lead to a call being set up. Also, the recognizer should work properly with any type of microphone, at a variety of distances and angles between the mouth and microphone, and despite changes from handset to hands-free equipment, all without having to retrain vocabulary.

The need for speech recognition

There are several reasons why speech recognition is becoming a standard feature in mobile phones. Using a mobile phone while driving a car has been regarded as dangerous, because it distracts the driver. During call set-up, the driver must remove his hand from the steering wheel to punch the telephone number into the keypad. To check the number, he has to take his eyes off the road and look at the display on the telephone. During the conversation, he needs to hold the telephone in his hand. Legal restrictions already exist or are being introduced (in Japan and Germany, for example), that prohibit a driver of a car to use a mobile phone while operating the vehicle unless the driver is using hands-free equipment.

The use of hands-free equipment is a step forward, as it allows the driver to keep his hands on the steering wheel during the conversation. During call set-up, speech recognition allows the user to speak the name of the person instead of keying in the telephone

Figure 1
Telephones using Ericsson's first-generation speech-control technology.



number. In addition, acoustic feedback of the recognized name can help to keep the driver's eyes on the road and hands on the steering wheel. Consequently, speech recognition can make it safer to use a mobile phone in a car.

Apart from gains in functionality, the development of mobile phones has been dominated by a decrease in the physical size of handsets in recent years. Increasingly smaller devices are being produced. In addition, customers are demanding bigger displays for new services. As a result, telephone keypads have diminished to a size that sometimes makes them awkward to operate. Every now and then, the news media feature "pen-like" phones with no keypad at all.

Speech is considered the most natural means of communication for humans. Thus, automatic speech recognition can become a natural user interface for mobile phones. It reduces manual interaction with mobile phones. In name dialing, for example, instead of searching the telephone book for a particular name, a user needs only speak the name, and the telephone automatically sets up the call.

New ways of using mobile phones are continuously being introduced. For example, Bluetooth headsets will allow the use of mobile phones with cordless portable hands-free equipment. These headsets and a speech-recognition interface offer a useful alternative to phone keypads. Other innovations will also heighten the need for reliable speech recognition.

Types of recognition system

The choice of the appropriate type of recognition system or recognizer for a mobile phone is crucial. Each type of recognizer has its own advantages and disadvantages.

Isolated-word vs. connected-word recognition

Isolated-word recognizers can recognize a single word in a recognition window. To have them recognize a sequence of words, the speaker must pause between each word to terminate and restart recognition, which results in unnatural pronunciation.

As the name implies, connected-word recognizers allow the speaker to say several keywords with no artificial pause between words. The cost is higher complexity compared to isolated-word recognizers. One problem for a connected-word recognizer is

coarticulation, the situation in which words are pronounced differently if spoken in a connected fashion, such as a sequence of digits. Word boundaries often disappear and words melt together. However, because of the natural style of speaking, connected-word recognizers are much more user-friendly than isolated-word recognizers and should thus be preferred.

Speaker-dependent vs. speaker-independent recognition

Users of mobile phones with current speech-recognition technology have to train their phones before they can use features such as name dialing. Training is seen as a cumbersome task for users. However, one advantage of such a speaker-dependent system is that it is language-independent. Any user can train his or her phone in any language desired. This simplifies the introduction of speech control as a new feature in a product such as a mobile phone, which is typically released worldwide in different countries where different languages are spoken. In addition, because the system learns each individual user's speaking behavior, performance is superior to that of speaker-independent recognizers.

Speaker-independent recognizers are pre-trained using a large sample of human speakers. A user can immediately start using such a system, which makes operation much easier. One drawback of such a system is that it is impossible to model all possible speaker variations for a language. Thus, there will always be a certain percentage of users whose phones will achieve sub-optimal performance. One solution to this problem would be to give users the option of training all or some words of the vocabulary that do not work well for them. This would require combining speaker-dependent and speaker-independent recognition in a single device.

Small-vocabulary vs. large-vocabulary recognition

For command and control applications, small-vocabulary recognition systems with vocabularies up to 100 words are sufficient. These recognizers are normally word-based, which means each vocabulary element consists of a word. The complexity and memory space required are limited.

In contrast, a large-vocabulary recognizer for the dictation of letters or e-mail can contain up to 100,000 words. The basic vocabulary elements are smaller sub-units of speech like phonemes, which are used to

BOX A, TERMS AND ABBREVIATIONS

Feature extraction

A preprocessing procedure in the speech-recognition process that transforms the speech signal into a spectral representation.

Feature vector

The result produced by feature extraction.

HMM

Hidden Markov model.

MIPS

Mega-instructions per second.

MMI

Man-machine interface.

Pattern matching

Also referred to as *search*—the procedure in speech recognition that matches the incoming utterance of the user against voice tags. The voice tags are also sometimes called *patterns*.

RAM

Random access memory.

SMS

Short message service.

Vocabulary

Collective term for all trained voice tags.

Voice tag

A representation of speech suited to a speech recognizer.

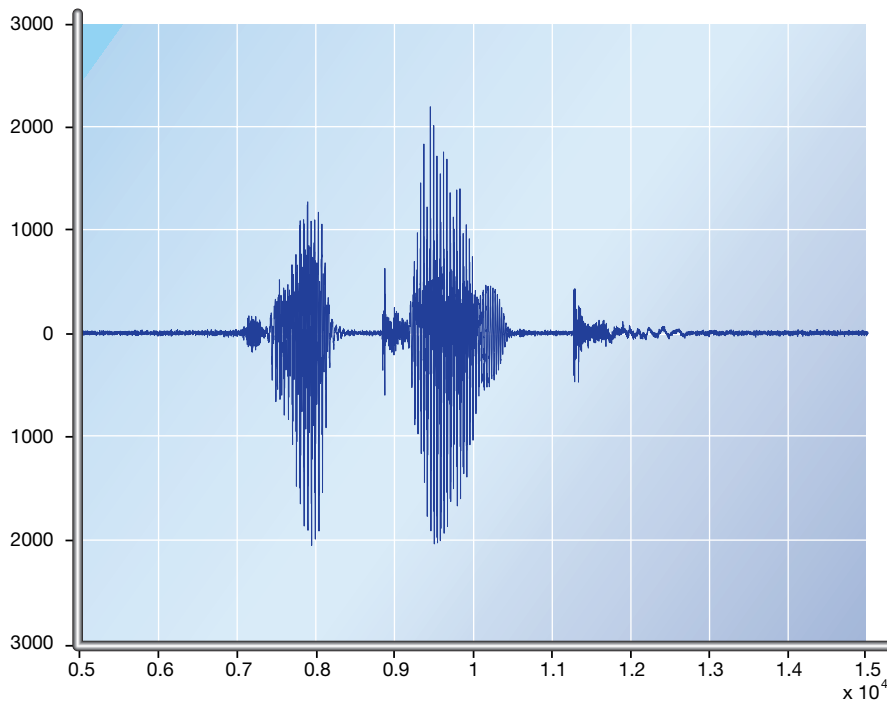
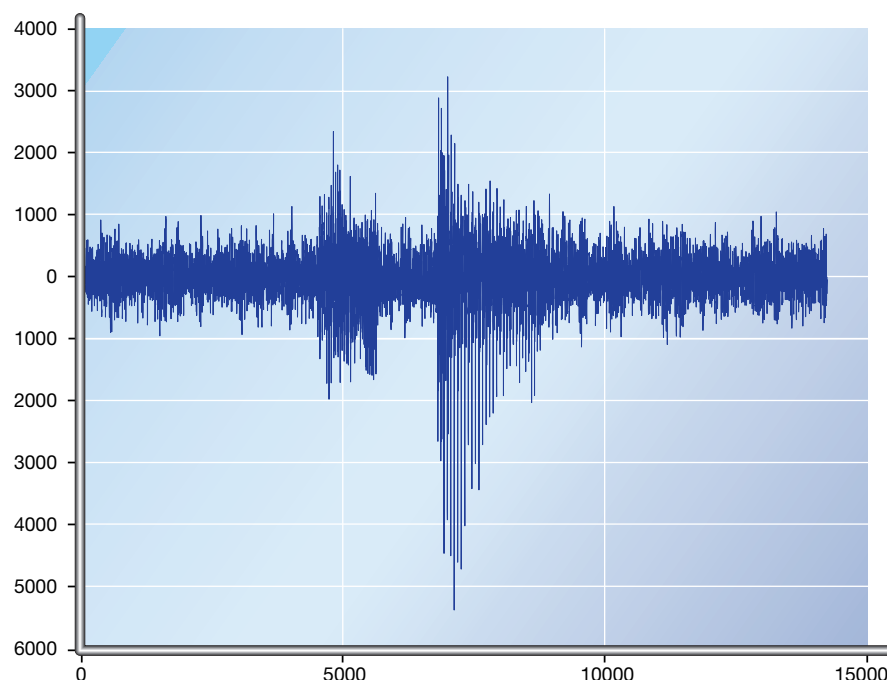


Figure 2
Without noise—speech file of the word *repeat*.

Figure 3
With stationary noise—speech file of the word *repeat*.



build up the words of the vocabulary. The complexity and memory space requirements are very high. Several dozen megabytes of memory are needed to implement a true dictation system, which is an unacceptable requirement for a mobile phone.

Challenges

Acoustics

Compared to mobile phones, fixed-line telephones typically offer a stable acoustic environment. They are used in the same room, such as an office or living room, with almost the same acoustic background every day.

In mobile communications, background noise is always present and extremely variable. Mobile devices are used in every imaginable environment. The setting could be an office or an airport, railway station, or even outdoors, with an acoustically challenging environment. Automotive interiors—with interfering background noise that depends on the type of car, the speed at which it is traveling, and so forth—are a challenge faced by mobile phones alone. Consequently, in mobile communications no assumptions can be made about the acoustic properties of the operating environment or background noise.

Also, a certain proportion of mobile users frequently change from handset to hands-free operation, either with built-in hands-free equipment in a car or with portable hands-free accessories. This causes large variations in the speech signal in addition to the conventional variation of attenuation from user to user.

Figures 2 - 4 show differences in speech signals caused by background noise in typical settings. The first illustration shows the speech signal of a word that has been spoken without background noise. Recognition of the spoken word should be a fairly simple task for any recognizer.

The second illustration shows the speech signal produced by the same word spoken with high stationary background noise. This situation corresponds roughly to hands-free operation in a car. The distance from the mouth to the microphone is about 30 centimeters. Background noise is caused by the engine, wind noise, and passing cars. A comparison of the speech signals in these two cases suggests problems of recognition. Increasing background noise degrades the performance of recognizers. There are techniques, such as spectral subtraction, for cop-

ing with noise in preprocessing or feature extraction as long as the noise is stationary.

A more extreme situation is shown in the third speech signal (Figure 4). Here, non-stationary noise comes from people talking in the background. This is the most demanding situation for an automatic speech recognizer, because methods dealing with stationary noise in feature extraction do not work here. The background noise is also human speech, so statistical methods do not help to distinguish between the desired input from the user and the undesired input from people talking in the background. Nevertheless, this is a rather typical situation for mobile phone use, and users expect their mobile phones to operate in all possible acoustic environments.

Hardware restrictions

Some characteristics of mobile phones influence the speech-recognition systems that can be used. In general, mobile handsets are

- very small devices;
- produced in large volumes; and
- highly cost-optimized.

This considerably restricts the hardware that runs the speech-recognition algorithms, as it limits capacity in terms of computational load, memory size, and memory access speed.

Computational load

To minimize the size and cost of handsets, no extra processor is allowed for the sole purpose of speech control. Existing computing resources have to be shared with other applications in the phone. Nevertheless, the digital signal processors used in mobile phones today offer performance of at least 40 mega-instructions per second (MIPS), up to 180 MIPS. Hence, this bottleneck is not a severe problem.

Memory size

The first generation of Ericsson phones with speech control had a vocabulary of 10 names. Ericsson Research has developed a solution that minimizes the additional memory required to store the voice tags and the auditory feedback of trained names. The problem of memory diminishes with modern system-programmable flash memory. This allows a relatively large vocabulary for name dialing and the implementation of new speech-controlled features in mobile phones. However, the memory available is still some orders of magnitude too small to allow the implementation of a large-

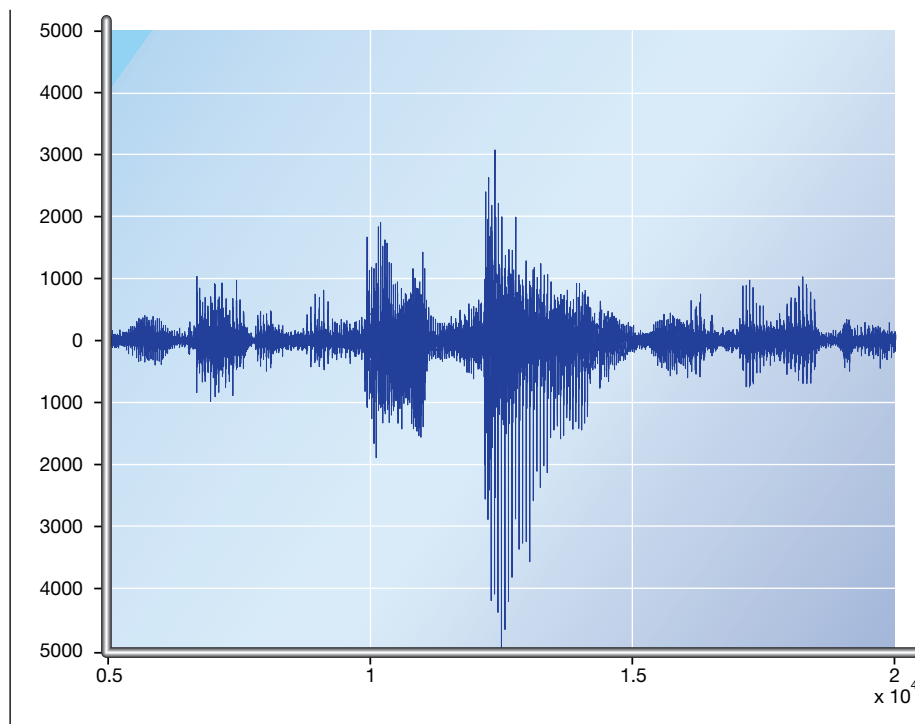


Figure 4
With non-stationary background noise—speech file of the word *repeat*.

vocabulary system for e-mail dictation and similar applications.

Memory access speed

Mobile phones typically consist of a multiprocessor solution with different memories attached to the available processors. The transfer of data between these memories tends to be a problem, because the normal hardware architectures of mobile phones were not designed for the rapid data-transfer rates needed for automatic speech recognition. The computational performance of the digital signal processors is good, but the performance of the processors' random access memory (RAM) is rather limited. Thus, access to the control processor flash memory is required.

Technology

The speech-recognition process

By definition, user-dependent speech-recognition systems require each user to run a training program to create the vocabulary for the recognizer. The program reads the

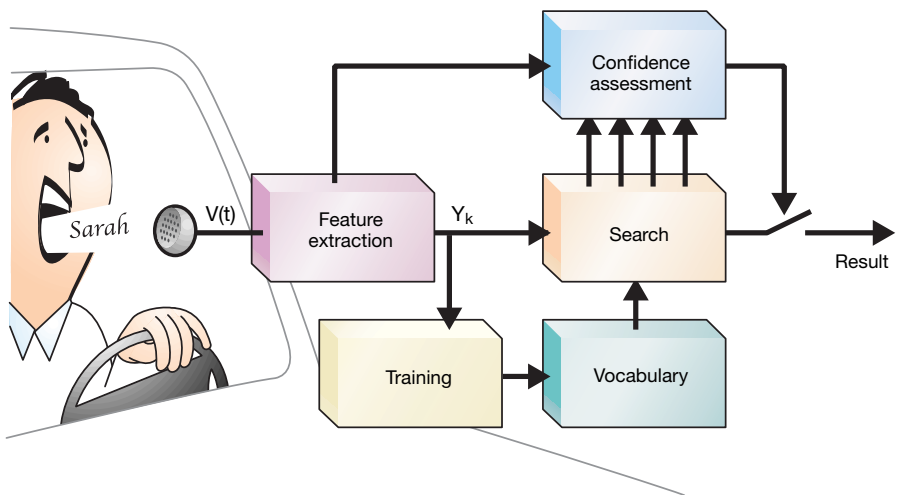


Figure 5
General block diagram of a speech recognizer.

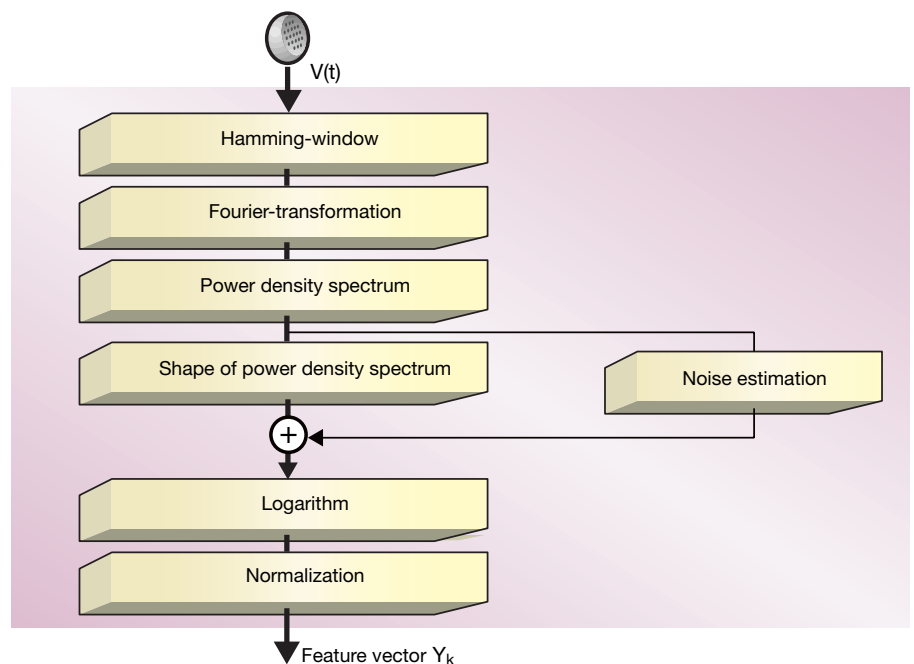
The speech signal $v(t)$ picked up by the microphone is fed into the feature-extraction function, which extracts essential information from the signal. The output of this function is the feature vector, y_k , which describes the basic acoustic properties of the speech signal for a given time interval. k denotes the index of a time interval. Typically, a new feature vector is generated every 10 ms.

The feature vectors are fed into the search function, which has additional input of the trained vocabulary—in the case of mobile phones, this normally consists of single-word models. The task of the search function is to match the sequence of incoming feature vectors, y_k , against the available vocabulary. The most likely vocabulary word becomes the recognition result. If a user unintentionally activates the recognizer, any speech signal activity could produce a result, thereby setting up a call. Obviously, this is unacceptable, so an additional block has been added: the confidence assessment function.

sequence of feature vectors that correspond to an utterance of the word to be trained and then creates and stores a word model of it. When the word is later used as a command, the system extracts features, searches to match the sequence of features in the utterance, and then runs confidence assessment (Figure 5).

After the search has finished a recognition pass, the confidence assessment function reads parameters from feature extraction (such as the signal-to-noise ratio during recognition) and search functions (such as the distance between first and second-best match). Based on these parameters, a deci-

Figure 6
Main blocks of the feature extraction function.



sion is made whether to accept and transfer the results to the overlying man-machine interface (MMI) or to reject the results of the search.

Feature extraction

Figure 6 shows the main blocks of the feature extraction function. The speech signal, $v(t)$, which consists of a continuous stream of speech samples, is partitioned for frame processing. Based on a sampling frequency of 8 kHz, typically 256 samples (that is, 32 ms window length) are combined to build a frame with an overlap of 176 samples over consecutive frames. The deployment of a Hamming window on each frame yields smoothing between frames. The Fourier transformation transforms the signal from the time domain into the frequency domain. This yields a feature vector that represents the basic spectral properties of the speech signal.

Subsequent processing steps use the power density spectrum instead of the complex Fourier spectrum. Investigations have found that human hearing is insensitive to phase variations, so the phase information of the signal is removed with this step. The fine structure of the power density spectrum carries extensive information on the speaker and his or her vocal characteristics. The information of the spoken utterance itself can be found in the envelope or shape of the spectrum. The envelope is derived by down-sampling the spectrum at distinct center frequencies, which are distributed in a nonlinear fashion over the frequency axis, again similar to human hearing. The Ericsson recognition system uses $\mu = 15$ center frequencies.

Stationary background noise during recognition is one important problem. Background noise can be modeled as an additive noise component N to the speech signal S in the power density domain. For each of the center frequencies, this can be described as:

$$X_{\Delta}(\mu) = S(\mu) + N(\mu)$$

The noise level N is estimated in the noise level estimation block. Subtracting the estimate of the background noise minimizes the influence of the noise signal on X_{Δ} , making the system more robust to stationary background noise.

Another problem is an acoustic environment with variations in microphone levels. One means of coping with this problem is

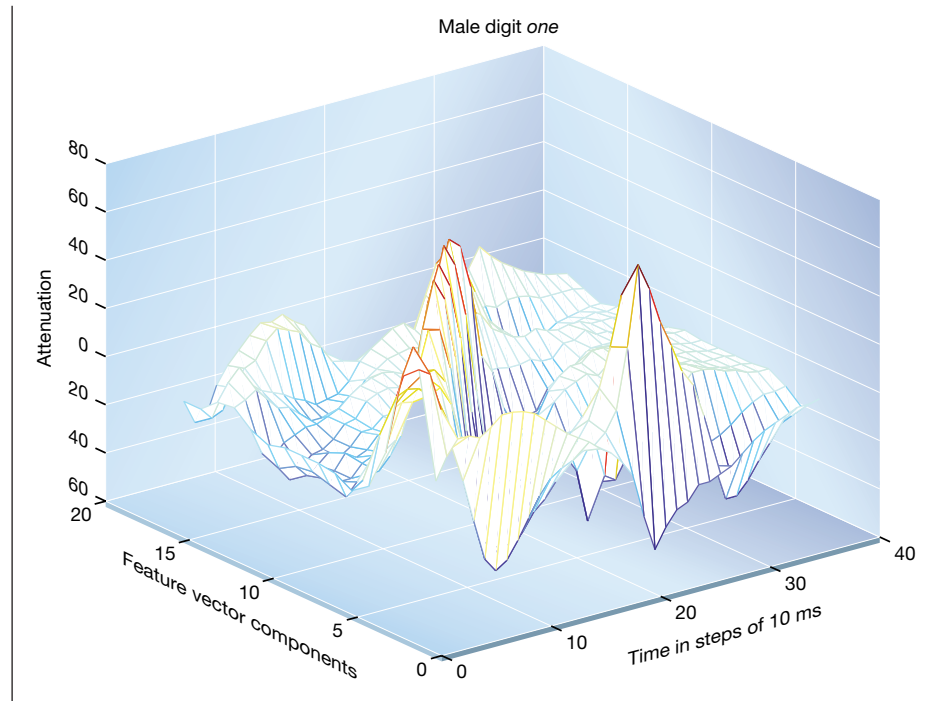


Figure 7
Sequence of feature vectors for the spoken English digit one.

to use the logarithm of the power density spectrum. $X_{\Delta}(\mu)$, which is the μ th center frequency before logarithmization, is multiplied by a constant factor, V .

$$VX_{\Delta}=10 \log\{V X_{\Delta}(\mu)\} \text{ with } \mu L \{l...15\}$$

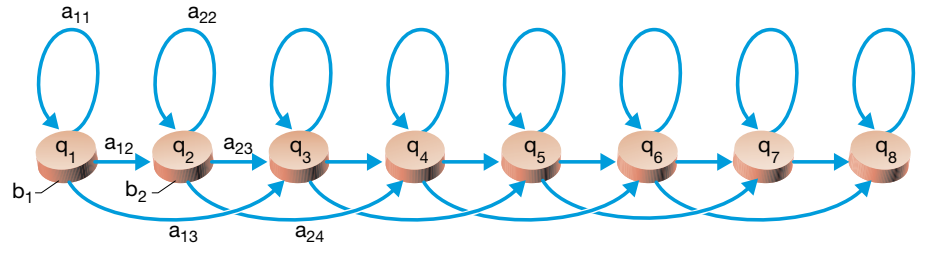
Bear in mind that for the Ericsson system, μ is between 1 and 15. The factor V stems from conditions such as one user's voice being louder than another user's voice. The above formula can thus be rewritten as:

$$VX_{\Delta}=10 \log\{V\} + 10 \log\{X_{\Delta}(\mu)\}$$

The constant attenuation factor V is decoupled in a sum from the desired signal $X_{\Delta}(\mu)$. Using differential parameters is an effective means of removing the additive term of the attenuation factor V . The mean energy of all center frequencies is subtracted and added as a 16th component. Further normalization steps lead to the final feature vector Y_k .

Figure 7 shows an example of the final result of feature extraction: the sequence of feature vectors for the spoken English word

Figure 8
Structure of a vocabulary word as a hidden Markov model.



one. The 16 feature-vector components are drawn over the sequence of 10 ms steps.

Search

As already mentioned, the search (or *pattern-matching*) function has the task of matching the sequence of incoming feature vectors from a user's utterance with the trained vocabulary of the recognizer. Ericsson's speech-recognition system employs hidden Markov models (HMM) to model human speech. With this technique, each word of the vocabulary consists of an HMM. The structure of an HMM (Figure 8) consists of a chain of states (denoted q), with each state describing a segment of the vocabulary

word. Thus, q_1 models the start, and q_8 models the end of the word.

States are connected with transitions, which facilitate state changes, depending on the transition probabilities, a_{ij} . Emission probabilities, b_i , which express the spectral similarity of a feature vector with a time interval of the reference, are attached to the states. Starting from the initial state, q_1 , different paths through the model can be used, depending on the sequence of incoming feature vectors. The repetition or skipping of states allows adaptation to variations in the rate of speech of the user. A word is recognized if a path through the reference has reached its final state, q_8 , with a reliable probability.

Figure 9
Cordless Bluetooth headset.



Future speech MMI for mobile phones

Experience gained from the first generation of voice control shows that the technology is robust enough to be used in mobile devices. A vocabulary of 10 names was sufficient as a starting point, but there is demand for larger name-dialing vocabularies. While the preferred size differs greatly among individuals, most users desire a larger vocabulary than currently available.

Up to now, the main application of speech recognition has been name dialing. Nevertheless, more functions in a mobile phone will be controlled by speech in the future. For example, a function can be directly activated with voice commands that might otherwise require a sequence of keypad entries in a layered menu interface structure.

The idea of speech shortcuts to menus is appealing, but more important is a speech MMI that does not leave the user dependent on visual feedback or complicated keypad interaction once he has used a voice command. A clear illustration of this issue is the use of mobile phones in a hands-free environment. In addition to current use in hands-free equipment in cars, hands-free operation will soon make use of cordless headsets based on Bluetooth technology (Figure 9).

The use of a cordless headset will allow a user to leave his or her telephone in a bag or out of reach. This requires control of at least the basic phone functions (accept or set-up a call, phone book administration) via the headset. The hands-free user will be dependent on audio feedback as a complement to speech input. Consequently, functions such

as call set-up, switching profile, and memo recording are suited to speech control, while functions such as viewing the calendar or short message service (SMS) inbox are not. Through careful selection of functions, the speech MMI will become invaluable to the mobile user.

The technology is still relatively elementary, owing to limited resources in mobile phones. Both computing power and memory size will increase in the future, allowing more sophisticated and user-friendly technology, such as connected-word recognition. It has also been suggested that speech recognition technology be used for applications requiring extensive vocabularies (more than 10,000 words), such as e-mail dictation. Although the resources in mobile phones will improve, the memory needed to host complex speech recognizers might never fit in a phone. However, one solution might be to combine terminal-based recognizers for frequently used telephone functions with complex network-based recognizers.

Conclusion

Despite the unique challenges that are posed by the mobile user's environment, robust speech-recognition systems are already used in many advanced handsets. Speech MMI will become increasingly important as mobile phone displays get bigger and their keypads smaller. Although hardware restrictions are likely to limit the size of on-board vocabularies for some time, users can look forward to functionality enhancements.