

The successful application of speech recognition in a product depends heavily on the specification of the product and the recognition technology used. This document provides an overview of the speech recognition technologies and design guidelines for increasing the reliability of speech recognition in consumer products. Effective use of the speech recognition technologies can not only enhance recognition accuracy, but also reduce memory requirements, and overall product cost.

Speech Recognition: Dependent vs. Independent

Through proper design and specification it is easy to create a product that is 99.5% accurate using the Interactive Speech™ technology. Determining the type of speech recognition to use is one key element in the product's specification.

The two general classes of speech recognition are "speaker-dependent" and "speaker-independent" recognition. With speaker-dependent recognition, the user trains the device to recognize his/her voice by speaking each of the words to be recognized several times. The product then recognizes these pre-trained words when spoken by the user. Training is a quick and simple process. In speaker-independent recognition, the product is *pre-trained* on the voices of *many* different speakers. The product is ready to use and requires no additional training by the user.

Speaker-dependent recognition yields slightly higher recognition rates than speaker-independent speech recognition because it is trained on a specific user's voice. For example, a speaker-dependent device recognizing the digits from 0 through 9 will have an accuracy higher than 99%, while a speaker-independent device performing the same task may have a recognition rate a few percentage points lower. Speaker-dependent recognition devices are ideal for applications requiring complicated recognition tasks. Speaker-dependent recognition systems will require storage of information about the speaker's voice; and thus, depending on the number of words and the Sensory processor chosen, these products may require some type of external memory (SRAM, EEPROM, or Flash).

Through proper product design, speaker-independent devices can yield high recognition rates. For example, a speaker-independent recognizer that distinguishes "yes" from "no" will have an accuracy higher than 99% because the recognition task is simple. More complicated recognition tasks can be performed with an accuracy just a few percentage points lower than speaker-dependent recognition if the product is designed carefully. Product design includes careful definition of the product's target users. This is extremely important in enhancing speaker-independent recognition rates. Speakers' accents, age, gender, as well as regional, socio-economic, and ethnic backgrounds are demographics that must be taken into consideration when *pre-training* the product for speaker-independent speech recognition. Because no additional writeable memory is required, this technology is ideal for inexpensive consumer electronic products. A final advantage of speaker-independent recognition is that it requires no training by the user - a product using speaker-independent recognition will work right out of the box.

The following table outlines the differences between independent and dependent recognition:

Feature	Speaker-Independent	Speaker-Dependent
Works right out of the box	Yes	No
Can recognize any language	No	Yes
Requires no writeable memory	Yes	No

The RSC-200/264T, RSC-300/364, and Voice Extreme™ offer options for both speaker-dependent and speaker-independent recognition.

Gameplay with Speech Recognition

In addition to technology selection, five keys to attaining the greatest recognition accuracy possible are:

1. Selection of an appropriate recognition vocabulary.
2. Questions which elicit appropriate answers.
3. Recognition triggers which activate the chip to “listen.”
4. Speech synthesis supporting the speech recognition tasks.
5. Control of background noise environments.

Below are some tips on scripting a product’s Gameplay for successful speech recognition.

Selecting Recognition Sets

The maximum number of SI or SD words in any given set depends on which Sensory IC is chosen. The success rate of recognition depends primarily on the following two parameters:

1. **Number of words in each set.** Every time a speech recognition product “hears” a word, it compares that word against those in the active recognition set. The more words a recognition set contains, the more likely the recognizer will make a mistake. It is important to limit the number of words in each recognition set whenever possible.
2. **Phonetic distinctiveness of each word in the recognition set.** Words in each recognition set must be chosen carefully. For example, a product that is required to distinguish between the words “three,” “free,” and “tree” will have a much lower recognition rate since these words are phonetically similar. Similarly, if the task is to distinguish “cat” from “rat,” the product will be improved if the recognition set is changed to “cat” and “mouse”.

At different stages during the product’s use, the recognizer can be programmed to listen for different sets of words. For example, a product that distinguishes the answers “yes” from “no” at one stage might later recognize a different word set which contains possible answers such as “dog,” “horse,” “elephant,” and “dinosaur.” The limited number of answer choices, the different numbers of syllables and the phonetic distinctiveness of each word combine to significantly enhance recognition accuracy.

Phrasing the Question

For the speech recognition chip to work successfully, a question must be phrased to elicit only one *specific* answer. The answer must be specific both by being uniquely correct (the only correct word) as well as by avoiding many forms. (E.g. a cloud, the cloud, cloud, the clouds, or clouds).

Uniquely Correct

Consider the question, “What is three plus four?” This is a good question because there is only one correct answer to this question. In contrast, the question, “Where does water come from?” is ambiguous and may elicit many different correct words, such as “clouds,” “the faucet,” “rivers,” “rain,” and “the reservoir.” If the product has been coded to accept only “clouds” as the correct answer, then it will appear not to work upon rejection of these other correct answers. A better question would be, “What is white and puffy and gives us rain?”; it is specific enough to rule out different vocabulary items like “the faucet,” “rivers,” “the reservoir,” etc. It is extremely difficult for the chip to choose intelligently from among all the possible correct answers to a loosely worded question.

Form

A question must also be crafted to elicit exactly one form of an answer. For example, the question, “What is white and puffy and gives us rain?” has a uniquely correct answer, but the correct answer could occur in different forms: “clouds,” “a cloud,” or “the clouds.” It is therefore important that questions be phrased such that they do not permit variations in number (singular vs. plural) or in use of an article (“a” or “the”). It would be an impractical goal for the recognizer to take into account the different combinations of number and article use for each answer.

Now consider the question, “Does water come from clouds or from sunshine?” From the standpoint of speech recognition accuracy, this is a good question because it elicits one of two specific answers, each with a single form (no variation of number or article use). Both possible answers, “clouds” and “sunshine,” would be

programmed into the recognizer and the product would be able to determine the correct answer with great accuracy.

Triggering the Speech Recognition Chip

There are three possible modes of operation for the speech recognition chip: normal, continuous listening and Wordspotting modes (not available on the RSC-200/264T). In continuous listening and Wordspotting modes, the chip is ready at all times to detect a recognition word or phrase. Wordspotting offers the ability to identify a long word or short phrase out of the middle of a sentence, whereas continuous listening requires that the trigger words be separated by silence. In normal mode, it is necessary to ensure that the chip knows when to listen. The latter is a more robust approach and should be used whenever possible; continuous listening and Wordspotting should be used only for specific applications that require this feature.

Ensuring the chip knows when to listen

The best way to ensure that the chip knows when to listen is through speech prompts (also referred to as speech synthesis). Usually, the end of a question (“What is three plus four?”) or a synthesis prompt (“Record or Play?”) helps to signal when the user is supposed to speak and when the speech recognition chip is expecting a response.

Another approach is to activate the chip manually by pushing a button before speaking. This approach can be avoided with appropriate use of speech synthesis.

Continuous Listening and Wordspotting

The use of continuous listening or Wordspotting will generally reduce the product’s overall recognition accuracy. It is difficult for the chip to clearly distinguish a single word or phrase from all the other audio signals that it receives when it is continuously listening. Random noise can be mistaken for the recognition word or phrase. Continuous listening can be implemented using either speaker-independent speech recognition or speaker verification; Wordspotting is speaker-dependent only.

Consider a wall clock that announces the time when you say a particular phrase. In a noisy environment such as an office or a conference room, it will occasionally announce the time even though the trigger phrase might not have been spoken. This apparent “mistake” occurs when the chip “hears” sounds that are phonetically similar to a trigger phrase.

The frequency of such “false positive” responses depends on the uniqueness of word to be recognized. For example, its operation may be acceptable if the word to be recognized includes a trigger recognition phrase. For example, the phrase “wall clock” can be used to trigger the clock to listen for the word “time”. If the word “time” is then said, the clock will know to announce the time. This two-step process will have a lower error rate than a recognition process only using the word “time.” Sensory recommends using two words for continuous listening to improve accuracy.

Power consumption is also an important factor in continuous listening. In operation, the speech recognition chip may draw a current of 10 milliamperes. If it is powered to continuously listen for some given phrase, it will drain a button battery in several hours or a large alkaline battery in several days. Thus, if the application requires that the recognizer be listening all of the time, it should operate from AC wall power. If the product is to operate on button batteries, then it must be awakened from a low power “sleep” mode for a period of a few seconds every time it is asked to recognize a phrase.

Synthesis Support

The Interactive Speech recognition chip can also interact with the user through speech synthesis to clarify responses. When recognizing words or phrases, Sensory’s recognition technology calculates its own probability of success. The product can thus be designed to prompt for cues if a desired level of recognition accuracy hasn’t been reached. For example, if a recognition doll is told to “walk” and the chip calculates that it has greater than an 80% chance of being accurate, it can accept the “walk” command. If it is only 50-80% sure of its accuracy, the recognizer can prompt, “Did you say walk?” The answer, “yes” or “no,” is very easy for the chip to determine and this provides a robust method of (eventually) getting the right answer. If less than 50% confident, the chip can ask, “What did you say?” effectively starting the recognition task over.

The following phrases should be included in all products using synthesis. These phrases will help improve overall product accuracy by notifying the user of possible problems:

- ▶ Please talk louder / Please talk into the microphone.
- ▶ Please talk softer.
- ▶ Please say...[each recognition item].
- ▶ Did you say... [each recognition item]?
- ▶ What did you say?
- ▶ You spoke too soon.

The following phrases may be desirable, depending on the product:

- ▶ I think you said...[*each recognition item*].
- ▶ In the picture, what do you see that...

Controlling the Noise Environment

Just like people, speech recognition systems have difficulty recognizing words in a noisy environment. Speech recognition products should be designed for use in a quiet environment whenever possible. If the product is meant to be used in a noisy environment, care must be taken to try to control the noise. For example, consider speech recognition being used in a video game with shoot-'em-up noise and music. Either the user should be provided with headphones for listening to the sounds (without their being heard by the microphone), or the sounds should be muted when the user is expected to talk, or a headset microphone should be used to provide a good voice signal to the recognition system.

Designing with Synthesis

Synthesis is the product's ability to "talk." The use of music, sound effects and speech synthesis can be easily incorporated to produce a product that is interactive and user-friendly. Speech synthesis allows the product to interact with user to provide feedback or directions as to how to better use the product. Although synthesis does require additional memory which can increase overall product cost, it's overall value to the end user is priceless.

Speech Synthesis Overview

Creating speech synthesis files for the Interactive Speech™ chips begins with making a list of all words/phrases the product will ever say. From this list, a recording script will be generated. If Sensory is performing the speech synthesis development, then the recording script and all subsequent steps will be done for you. If not, then the next step is to use a voice talent to record phrases based on the script. After recording, the best of all the recorded files must be chosen for use in the product. These files are then compressed into one master file (a .O file) that a developer will use to make the chip talk. This can be done either by Sensory, or by the developer using Sensory's Quick Synthesis application. Creating speech synthesis using Quick Synthesis can be done in minutes, but with some sacrifice in speech quality and data compression compared with Sensory-compressed files.

A technical note: The term "speech synthesis" can be misleading. The Interactive Speech™ chips perform "time domain compression" that sometimes re-uses sections of speech in different words. This re-use of speech sections is the reason this process is referred to as "synthesis." In general, however, it is more useful to think of the Interactive Speech™ line as using "time domain compression" instead of "speech synthesis".

Re-using Words and Phrases

Overall, memory requirements for synthesis can be reduced by re-using words and phrases as much as possible. However, the re-use of words and phrases is not as simple as it may appear. A given word is not always suitable for re-use in other phrases. Words cannot be strung together in any order to form any sentence. Words contain qualities such as intonation, volume, duration, and emphasis that can vary according to the meaning and function of the sentence in which they occur. These qualities must be preserved in order to replicate natural-sounding speech. The challenge for a scriptwriter is to balance these considerations and maximize the vocabulary within a limited amount of memory.

Pronunciation of a given word often depends on context. For example, the word “the” is pronounced several different ways depending on the context. “The” is usually pronounced “thee” before vowels (e.g., “thee elephants,” “thee apple”). Before consonants, “the” is often pronounced as “thuh” or just “th” (e.g., “thuh red,” “th fish”).

Intonation is also an important factor in determining the suitability of words and phrases for re-use elsewhere in the product. By “intonation,” we mean the melody or variations of pitch in spoken words and phrases. The intonation of a word or phrase depends on its location in a sentence, as well as the meaning of the sentence. Proper intonation not only sounds natural but also provides listeners with important cues revealing the speaker’s meaning and intent. For example, consider the word “fish” in “Is this a fish?” and “This is a fish.” Different versions of “fish” should be used in these sentences to ensure natural-sounding, comprehensible speech. Recording a script in a near monotone would reduce variations in intonation and allow greater re-use of words. However, the result would be dull, stilted speech.

Re-use of speech is thus more effective if a script preserves sentence patterns and re-uses sentence units. If the number of sentence patterns is kept to a minimum, units of speech can be more readily interchanged. For example, consider the sentences “At the beep, press a button,” and “At the beep, tell me your name.” The single phrase “at the beep” can be used in both sentences because they call for identical intonation and emphasis. This phrase could not be re-used in a sentence with a different pattern, such as “Say your name at the beep,” without losing the intonation that makes speech sound natural. Given the proper conditions, words and short phrases can be used in different parts of sentences, or even in different sentence structures. The more the speech is chopped up and recombined, however, the more disconnected the resulting compressed speech will sound.

Counting Words in a Script

It is essential to estimate how much memory will be required and what level of effort/cost will be associated with compressing the speech synthesis phrases. These factors affect the overall cost of the product and its development. Counting words is the best way to approximate the total requirements. It is usually best to count words conservatively (e.g., count “fireman” and “firehouse” as two words each).

Each time a word is said in a different sentence, in a different way, or surrounded by different words, that word should be counted again. For a conservative estimate, it is best to repeat phrases, not words. Sensory’s team of expert linguists can evaluate phrase lists and determine how words or phrases may be formulated for re-use within the product.

Synthesis in More Detail

This section details exactly how much sound (words, character-voices, sound effects, music, beeps, and bells) can fit on an RSC-264T or RSC-364. The total amount depends on both subjective (e.g. sound quality) and qualitative (e.g. compression rate) factors.

The procedure for determining the number of seconds of sound playback available: Calculate the amount of **available ROM** (in bits) and **divide** by the **data rate** (bits/second) for the sound you want played back.

Available ROM

The RSC-264T and RSC-364 chips have 64 kilobytes (that’s 64,536 bytes or $64,536 \times 8 = 524,288$ bits) of ROM on-chip – but not all of it is available for sound output. The ROM space is used for other functions, like speech recognition. Ultimately, even a product with an emphasis on speech output will have under 50kB (409,600 bits) of ROM available on-chip for the customer to store sound output. The customer can add as much off-chip ROM or Flash as desired in order to increase the sound capability of the product. External ROM or Flash is required with the RSC-200 and RSC-300. Voice Extreme™ is designed to operate in conjunction with a 2MB Flash memory device.

Data Rate

Sound exists as an analog signal that must be converted to a digital signal for storage on a chip. During this conversion, recorded sound is compressed to reduce the amount of storage needed on-chip. The RSC IC can compress sound down to data rates from 15,000 bits/sec to under 5,000 bits/sec. The data rate depends on the following factors:

1. **Desired quality.** Sound compression involves a tradeoff between quality and quantity. The more we compress the sound, the more will fit on the chip. The less we compress it, the better it sounds. This is subjective; what sounds great to a toy designer may sound horrible to an audiophile.
2. **Nature of the sound.** Some sounds compress better than others. If a customer *wanted* the product to sound like a robot, then we could achieve very low data rates (under 5,000 bits/sec). However, a high-pitched voice, such as a licensed reproduction of Minnie Mouse, might require more than 15,000 bits/sec. There are two reasons for this. First, we want to retain the *personality* of Minnie Mouse’s voice. Second, high-pitched sounds don’t compress as well as low-pitched ones. As another example, sound effects are often very difficult to compress, requiring up to 20,000 bits/sec.
3. **Repeatability of the sound.** After you use a sound once, you can use it again without requiring extra memory. This can happen on several different levels. Sound effects, for example, can be looped such that a single “whomp” plays repeatedly to produce the sound of a helicopter: “whomp whomp whomp whomp...” Words and phrases can also be re-used in the same part of different sentences or, sometimes, in different parts of sentences. Our ability to do this depends on several factors: the desired quality, the script, and the intonation of the words and phrases selected for use in synthesis.

As a baseline, a custom masked RSC can fit from 25 to over 100 seconds of sound on-chip. A word usually takes about 0.5 sec and a sound effect about 1sec. Therefore, the RSC chips can fit from approximately 50 to 200 words on-chip. As mentioned previously, additional sound can be stored in off-chip ROM or Flash.

Sensory Support Services

Sensory offers a variety of services and resources to help product designers integrate speech into their products.

Speaker-Independent Recognition Sets

Currently it is required that Sensory creates the independent recognition sets for use with the Interactive Speech™ chips for all customers. This section identifies the information and steps required for building speaker-independent sets.

The following information is required to create the sets:

1. List of recognition words divided into appropriate recognition sets.
2. Description of the environment the product will be used in.
3. Recordings of approximately 500 people saying the entire word list. The demographics of this pool must accurately reflect that of the product’s target market.
4. Sample background noises.

The recognition list (include English translation if appropriate) should be arranged with the words in a single list as well as in their respective sets. The following is an example of a recognition list for a simple record and playback device:

Set 1	Set 2	Entire List
1. Record	1. Yes	Record
2. Play	2. No	Play
3. Erase		Erase
		Yes
		No

Sensory will review the word list, paying careful attention to the selections in each set in order to ensure accurate recognition. If Sensory does not foresee any problems, then the acquisition/recordings of the appropriate words can start.

To record voices, an IBM PC or compatible computer with a high quality sound card is required. Sensory will support the customer or sales rep as needed.

Speech Synthesis

Sensory has available services that allow them to add speech synthesis to their products. Current procedure requires that Sensory create the speech synthesis files for use with the Interactive Speech™ chips. In order for Sensory to create synthesis files, the customer should create a phrase list containing every phrase to be spoken by the product.

The phrase list should contain all phrases to be said by the product. Include a translation into English if appropriate. No phrase or word concatenation should be performed at this stage. An example of a phrase list for a simple record and playback device follows:

- | | |
|----------------------------|--|
| 1. Record, play, or erase? | 8. Just say: record, play, or erase. |
| 2. Recording. | 9. Did you say "record"? |
| 3. Playback. | 10. Did you say "play"? |
| 4. Erased. | 11. Did you say "erase"? |
| 5. Please talk louder. | 12. Please wait until I'm finished / Please wait for the beep. |
| 6. Please talk softer. | 13. What did you say? |
| 7. It is too noisy here. | |

A development station is being designed that will enable customers to create their own synthesis files. Contact Sensory for further information.

Music Synthesis

Sensory also provides customers with services that allow them to use music synthesis in their products. Currently Sensory creates the music synthesis files for use with the Interactive Speech™ chips for customers. In order for Sensory to compress musical compositions for use on the Interactive Speech™ chips, the customer must create files in four-voice MIDI format. In these MIDI files only certain percussive and musical instruments, with designated note ranges, are compatible with Sensory's compressed music note library. These restrictions on the MIDI instruments and note ranges are specified in the following table:

Acceptable Instruments and Note Ranges	
Piano	D2-E7
Trombone	C3-C6
Clarinet	C3-E7
Banjo	C3-E7
Bass Drum	n/a
Snare Drum	n/a
Cowbell	n/a

In the above notation system, C3 is two octaves below C5, which designates middle C (C5 = 523.25Hz).

No more than four 'events', i.e. any combination of four musical notes and/or drumbeats, may play at any one time for music files. For example, at time 04:01:000 in the composition only one piano note, one trombone note, one banjo note and one percussion beat may play simultaneously. Likewise, a piano chord of, say, three notes may be accompanied by only one other event, such as a single clarinet note.

For the fullest sound, the composer should make the most of the four voices available. If you do not have in-house music talent that can recompose songs to meet the above criteria, you can send a "plain" MIDI file to Sensory and we will recompose it for you. The only disadvantage to this approach is that the song may sound slightly different than you expect by the time we have made adjustments for compatibility. To ensure the best reproduction, we recommend sending an audio tape.

Other Support Services

Product design, product definition and other support services are available from Sensory by arrangement.

The Interactive Speech™ Product Line

The Interactive Speech line of ICs and software was developed to “bring life to products” through advanced speech recognition and audio technology. The Interactive Speech Product Line was designed for consumer telephony products and cost-sensitive consumer electronic applications such as home electronics, personal security, and personal communication. The product line includes award-winning RSC-series general-purpose microcontrollers and tools plus a line of easy-to-implement chips that can be pin-configured or controlled by an external host microcontroller. Sensory's software technologies run on a variety of microcontrollers and DSPs.

RSC Microcontrollers and Tools

The RSC family of microcontroller (RSC-300/364) are low-cost 8-bit microcontrollers designed for use in consumer electronics. All members of the RSC family are fully integrated and include A/D, pre-amplifier, D/A, ROM (RSC-364), and RAM circuitry. The RSC family can perform a full range of speech/audio functions including speech recognition, speaker verification, speech and music synthesis, and voice record/playback. The family is supported with a complete suite of tools and development kits.



Application Specific Standard Products (ASSPs)

- **Voice Direct™ 364** provides inexpensive speaker-dependent speech recognition and speech synthesis. This easy-to-use, pin-configurable chip requires no custom programming and can recognize up to 60 trained words in slave mode, and 15 words in stand-alone mode. Ideal for speaker-dependent command and control of household consumer products, Voice Direct* 364 is part of a complete product line that includes the IC, module, and Voice Direct 364 Speech Recognition Kit.
- **Voice Extreme™** simplifies the creation of fully custom speech-enabled products by offering developers the capability of programming the chip in a high-level C-like language. Program code, speech data, and even record and playback information can be stored on a single off-chip Flash memory. Based on Sensory's RSC-364 speech processor, Voice Extreme includes a highly efficient on-chip code interpreter, and is supported by a comprehensive suite of low-cost development tools.



Software and Technology

- **Voice Activation™** micro footprint software provides advanced speech technology on a variety of microcontroller and DSP platforms. A flexible design with a broad range of technologies allows manufacturers to easily integrate speech functionality into consumer electronic products.
- **Fluent Speech™** small footprint software recognizes up to 50,000 words; offers Animated Speech with the ability to automate enunciation and articulation; performs text-to-speech synthesis in either male or female voices; provides noise and echo cancellation, performs Wordspotting for natural language usage; offers telephone barge-in; and provides continuous digit recognition.



Important notices

Reasonable efforts have been made to verify the accuracy of information contained herein, however no guarantee can be made of accuracy or applicability. Sensory reserves the right to change any specification or description contained herein. Sensory reserves the right to make changes to or to discontinue any product or service identified in this publication at any time without notice in order to improve design and supply the best possible product. Sensory does not assume responsibility for use of any circuitry other than circuitry entirely embodied in a Sensory product. Information contained herein is provided gratuitously and without liability to any user. Reasonable efforts have been made to verify the accuracy of this information but no guarantee whatsoever is given as to the accuracy or as to its applicability to particular uses. Applications described in this data sheet are for illustrative purposes only, and Sensory makes no warranties or representations that the RSC series of products will be suitable for such applications. In every instance, it must be the responsibility of the user to determine the suitability of the products for each application. Sensory products are not authorized for use as critical components in life support devices or systems. Sensory conveys no license or title, either expressed or implied, under any patent, copyright, or mask work right to the RSC series of products, and Sensory makes balance between recognition and synthesis no warranties or representations that the RSC series of products are free from patent, copyright, or mask work right infringement, unless otherwise specified. Nothing contained herein shall be construed as a recommendation to use any product in violation of existing patents or other rights of third parties. The sale of any Sensory product is subject to all Sensory Terms and Conditions of Sales and Sales Policies.



1991 Russell Ave., Santa Clara, CA 95054
Tel: (408) 327-9000 Fax: (408) 727-4748

© 2001 SENSORY, INC. ALL RIGHT RESERVED.
Sensory is registered by the U.S. Patent and Trademark Office.

All other trademarks or registered trademarks are the property of their respective owners.