MASTER

Speech recognition for an environmental control system

Custers, R.J.G.

Award date:
1997

Faculteit der Elektrotechniek
Technische Universiteit Eindhoven
Vakgroep Meet- en Besturingssystemen (MBS)
Sectie Medische Elektrotechniek

Speech recognition for an environmental control system

R.J.G. Custers

Rapport van het afstudeerwerk
uitgevoerd van augustus 1996 t/m mei 1997
in opdracht van plaatsvervangend leerstoelhouder Dr. Ir. P.J.M. Cluitmans
onder begeleiding van Ir. W.H. Leliveld en dhr. H.J.M. Ossevoort

De faculteit Elektrotechniek van de Technische Universiteit Eindhoven aanvaardt geen verantwoordelijkheid voor de inhoud van stage- en afstudeerverslagen.

Preface

This report contains the results of my graduation project carried out at the section Medical Electrical Engineering in the Department of Measurement and Control Systems of the Faculty of Electrical Engineering at the Eindhoven University of Technology (EUT). The graduation project was performed during a ten month period (August 1996 - May 1997) to form the completion of the five year curriculum Information Technology.

I would like to thank Dr. Ir. P.J.M. Cluitmans for giving me the opportunity to do my graduation project in the section Medical Electrical Engineering, my coaches Ir. W.H. Leliveld and Mr. H.J.M. Ossevoort for their guidance, and last but not least my fellow students, and the other members of the section who made my project not only instructive, but also very pleasant.

Summary

This report covers a feasibility study to employ present speech recognition technology for environmental control by motor disabled persons. The speech recognition system serves as an input device to operate the X-10 environmental control system by voice commands. The speech driven interface, should be small, stand-alone, low cost, and has to perform speaker dependent isolated word recognition for approximately 30 words.

Considering the costs, the speech driven interface can best be built using standard components. An extensive field study on commercial speech recognition IC's led to five potential candidates. The MSM6679 voice recognition processor from Oki Semiconductors meets the requirements for the speech recognition system, and is selected.

Several tests on speaker dependent isolated word recognition were conducted, to obtain an honest indication of the MSM6679's recognition performance. The results were encouraging (%correct > 92%), and the project was continued with the MSM6679.

The recognition performance of the complete system can be increased with the use of a sophisticated user interface. A literature study to user interface design aspects for voice controlled devices resulted in many useful proposals regarding: feedback, vocabulary management, microphone placement and training of the system. These human factor proposals are incorporated in the design, and together with the feedback features of the MSM6679, they provide for a sophisticated command dialog.

The implementation comprises a micro-controller (μC) to command the host driven speech recognition processor to perform the various recognition and synthesis tasks. Next to this, the μC performs the high level operations to support the user interface and vocabulary management.

Since the communication of the X-10 system is performed over the mains, there is a substantial danger for leakage currents. To avoid any physical contact with the mains, the X-10 system is operated by infra-red signals. The infra-red codes are composed in software on the μC, and sent by an infra-red transmitter circuitry. Because of that, the device can be used to generate infra-red signals for operation of other appliances such as TV's, and radio's, as well. This extends the input device for the X-10 system to a flexible speech driven remote control.

The eventual experimental prototype can wirelessly operate X-10 appliances, by voice commands. Therefore, it exploits a sophisticated command dialog with extensive auditive feedback. Several human factor design proposals are incorporated in the prototype. Nevertheless, as functionality is the most important criterion for the design in this first stage of the project, the implementation is not yet optimized for energy consumption, size and user friendliness. In future, further work has to be done in optimising the system.

Contents

1. Introduction

Speech is a natural communication means for humans; speech is familiar and convenient. For some people, speech may even be one of the few options available, allowing them to communicate with other persons, and to have some degree of control over their daily environment. A speech driven environmental control system is far more friendly for disabled persons, than the usage of pneumatically or eye-movement driven switches. For these reasons, speech seems to be a good means, for disabled persons, to control their activities in daily living by spoken commands.

For quite some years, the project: instrumentation technology for elderly and disabled persons[1] in the Medical Electrical Engineering group has been developing customized remote control systems, for (motor) disabled persons. In the line of the "Monoselector project"[2], a study was started in 1984 for motor disabled persons, to control their environment by means of speech commands.

The speech recognition system was used as a customized interface to a commercial environmental control unit: Busch Timac X-10. The X-10 system can control (i.e. switch on/off and brighten or dim) up to 256 devices plugged into mains sockets, throughout the house. The X-10 system consists of a central control panel, and terminal plug-in units in different places in the house. From the control panel, the domestical appliances, that are powered through the terminal plug-in units, can be operated. In this way, one can operate lights, doors, curtains, etc., throughout the house, from one central control panel. The control center and the plug-in units of the X-10 system communicate over the existing (in house) mains, so no additional wiring is required. Nowadays, the X-10 system is becoming the defacto standard for environmental control. Extensive information on the X-10 system is described in [Bloks87] and [Krol90]. An ongoing discussion on the X-10 system can be found in the X-10 FAQ on the World Wide Web.[3]

The speech interface to the (control center) of the X-10 system was a subject of study for more than seven years. It was initiated in 1984 by H. Bosch [Bosch85]; with an extensive study on the possibilities of a (low cost) speech driven environmental control system, with the speech technology at that time. The speech recognition and synthesis chip: SP1000 from General Instrument, was chosen to be used as an isolated word speech recognizer. Bosch set up an experimentation system, to examine the device thoroughly. During the project, numerous improvements were made to increase the recognition performances of the system. Nevertheless, due to persistent disappointing recognition results, the project was stopped.

[1] This project is concerned with the development of new equipment or adjustment of existing apparatus for convenient use by elderly or disabled persons.

[2] This is a remote control system, for motor disabled persons. With this device linked to an environmental control system, a disabled person can control part of their immediate domestic environment, using only one button (switch).

[3] URL: http://www.homation.com/x10faq

Now, more than ten years ahead of SP1000's technology, the study is reopened. We expect that today's speech technology can provide a low cost speech recognizer, with adequate recognition performance for usage in a speech driven environmental control system, based on the X-10 system. The goal of this project is to examine the possibilities (and performance) of nowadays speech technology, and to implement a low cost, stand alone speech recognition interface to the X-10 system, for (motor) disabled persons.

2. Automatic speech recognition.

This chapter gives a brief overview of the different types of, and approaches to automatic speech recognition. First the different types of speech recognition are shown, next the different approaches to speech recognition are shortly described, and the pattern recognition method is wider discussed.

2.1 Types of speech recognition

In this paragraph different types of speech recognition systems are discussed in a nutshell, to place the kind of speech recognition used in the environmental control appliance.

The two general classes of speech recognition are:

1. isolated word (discrete) speech recognition,
2. continuous (connected) speech recognition.

Isolated word speech recognition requires the user to pause before and after utterances that are to be recognized. This is the simplest form of recognition, as the word boundaries are easy to find, and the different (isolated) words do not tend to affect each other. A continuous speech recognition system operates on speech in which words are connected together, i.e. not separated by pauses.

Continuous speech is more difficult to handle because of a variety of effects. First, it is difficult to find the start and end points (boundaries) of words. Another problem is "coarticulation"; the sound of each word (and even syllable) is affected by the surrounding words in the continuous speech stream. Similarly the start and end of words are affected by the preceding and following words. The recognition of continuous speech is also affected by the rate of speech.

A second distinction in speech recognition systems can be made on the basis of the number of different voices (users) that should be recognized:

A. speaker-dependent (SD) speech recognition,
B. speaker-independent (SI) speech recognition.

For speaker-dependent speech recognition the speaker trains the systems to recognize his/her voice by speaking each of the words to be recognized several times. In speaker-independent recognition the recognizer is not trained by one specific voice, but with a large set of samples from many different speakers. In this way the speech recognizer can recognize many different speakers. Because of differences between speakers, speaker-dependent recognition always yields better results, than speaker-independent speech recognition does.

A speaker-dependent isolated word recognizer is the simplest and cheapest recognition system of all combinations. Nevertheless, it is suitable for voice commands in environmental control (see chapter 3 on user-interface design).

2.2 Principal architecture of a speech recognition system

All speech recognition systems are globally constituted of the same blocks. Figure 2-1 shows a general overview of the composition of a speech recognition system. The various blocks are discussed below.

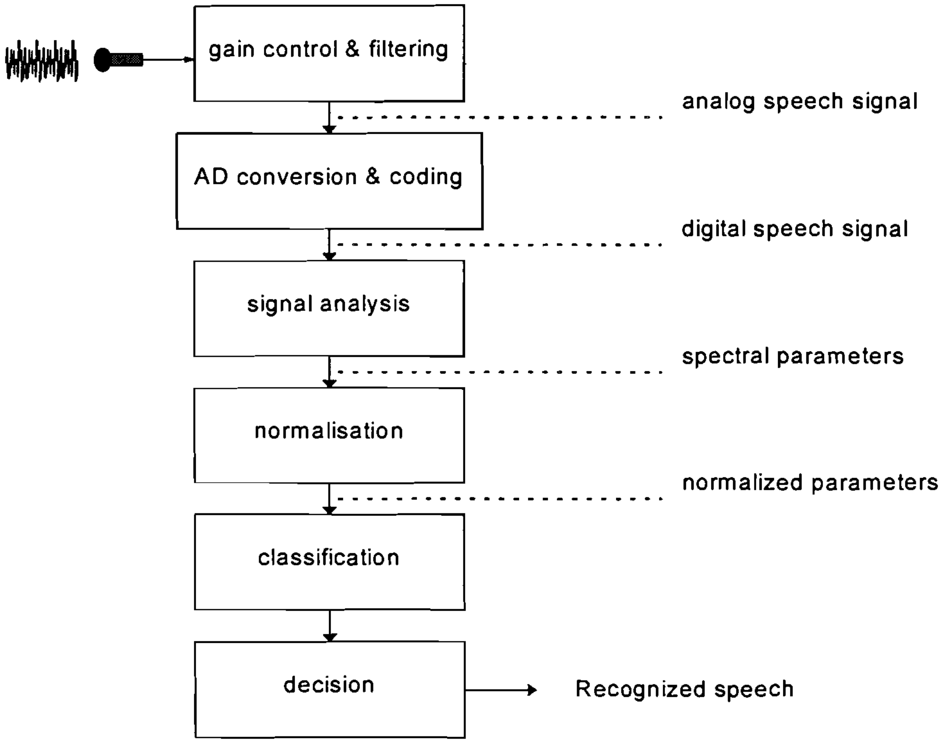


Figure 2-1, Global block diagram of a speech recognition system.

Input stage.

The first stage of a speech recogniser is of course the microphone; it converts the acoustic speech signal to an electrical one. In the next step low pass filtering is applied, as the speech spectrum of 100-8000 Hz spans the frequency range of interest. The (automatic) gain control is applied to compensate for different speaking volume and background noise level. The next phase contains the digitizing of the analog speech signal (AD-conversion) and some kind of coding of the digitized samples, to get a convenient digital representation of the speech signal.

Signal analysis.
The digital (coded) speech-signal is now led through the signal analysis box; the goal of speech analysis is to provide a compact (spectral) representation of the characteristics of the time-varying speech signal. This representation is often expressed as feature vectors. The short time power spectrum of speech is still considered to be the most effective representation for speech recognition [Atal95]; the power spectrum can be obtained from a filterbank, Fourier transforms, or linear prediction analysis (LPC). In [Rabiner93] a nice overview of signal processing and analysis methods for speech recognition can be found.

Normalization.
After signal analysis, a set of parameters describing the (power spectrum) properties of a time frame of approximately 20 milliseconds of speech signal is available. To compensate for variations in speech rate and volume, linear normalization steps can be performed. For a discussion on time- and amplitude- normalization see [Poll91].

Classification and decision.
All operations performed on the speech signal until now are rather common to all speech recognition systems; all those steps are performed to get a convenient, normalized representation of the speech signal. The next part in the speech recognition process comprises the actual recognition of the speech. The classification process classifies the speech signal of a captured utterance to a certain reference pattern (i.e. word, syllable, phoneme, etc.). The decision block at the end of the speech recognition process decides whether the speech signal is "close enough" to the pattern it was classified to, and leads to the recognized speech.

The classification can be done in a lot of ways; the classification (and decision) actually is the part at which the various speech recognition techniques differ from each other. In the next paragraph three different approaches to this classification process are discussed.

2.3 Approaches to automatic speech recognition.

In this paragraph three approaches are discussed in the way they differ in the "classification" part of a speech recognition system. The pattern recognition approach (most used) is wider discussed.

Broadly speaking, there are three approaches to automatic speech recognition [Rabiner93]:

1. the acoustic phonetic approach,
2. the pattern recognition approach,
3. the artificial intelligence approach.

The **acoustic phonetic approach** is based on the theory of acoustic phonetics; this theory is based on the principle that there is a finite number of distinct phonetic units in spoken language. The phonetic units are called "phonemes"; phonemes are the smallest **acoustical** components of a language. The idea behind the acoustic phonetic approach is to recognize these phonemes and build up a possible word or sentence out of the separate phonemes. If the constructed word is in the vocabulary of the speech recognition system, then it is recognized.

Contrary to the acoustic phonetic approach, the **pattern recognition method** for recognizing speech does not use any knowledge whatsoever about speech or language. The pattern recognition technique gathers its "knowledge", to recognize speech, by being shown examples of speech patterns. This "training" of patterns is used in most pattern recognition techniques (not only speech pattern recognition).

The **artificial intelligence (AI) approach** makes use of both the pattern recognition approach and the acoustic phonetic features of speech. This approach makes use of several "knowledge sources" to "understand" the speech signal. This includes acoustic and phonetic knowledge (like in the acoustic phonetic approach), lexical knowledge (concerning the words in the language), syntactic knowledge (the grammar of the language), semantic knowledge (concerning the meaning) and even pragmatic knowledge (about the application domain). Some of the techniques used in the AI-approach include expertsystems and neural nets [Rabiner93].

The most used technique nowadays is the pattern recognition method, as it is simple to implement with the current state of statistical and mathematical methods. The recognition performance is high (in speaker-dependent isolated word recognition), and robust. In the next paragraph the pattern recognition method is wider discussed.

2.4 The pattern recognition approach.

In this paragraph the most used speech recognition technique, called pattern recognition, is discussed; first the architecture similar to all pattern recognition techniques to perform speech recognition is discussed. Next, the template matching technique, and a statistical technique using hidden Markov models are wider discussed.

2.4.1 General pattern recognition.

Inherent to pattern recognition techniques is the fact that before recognition can be performed, the patterns must be trained. Speech recognition systems based on pattern recognition techniques employ two modes of operation (see figure 2-2):

1. a training mode, in which speech patterns of known classification are stored as reference templates (i.e. a fingerprint from the speech pattern),
2. a recognition mode, where an unknown incoming speech pattern is compared with the reference templates, and classified to one of the stored templates based on "some means of similarity".

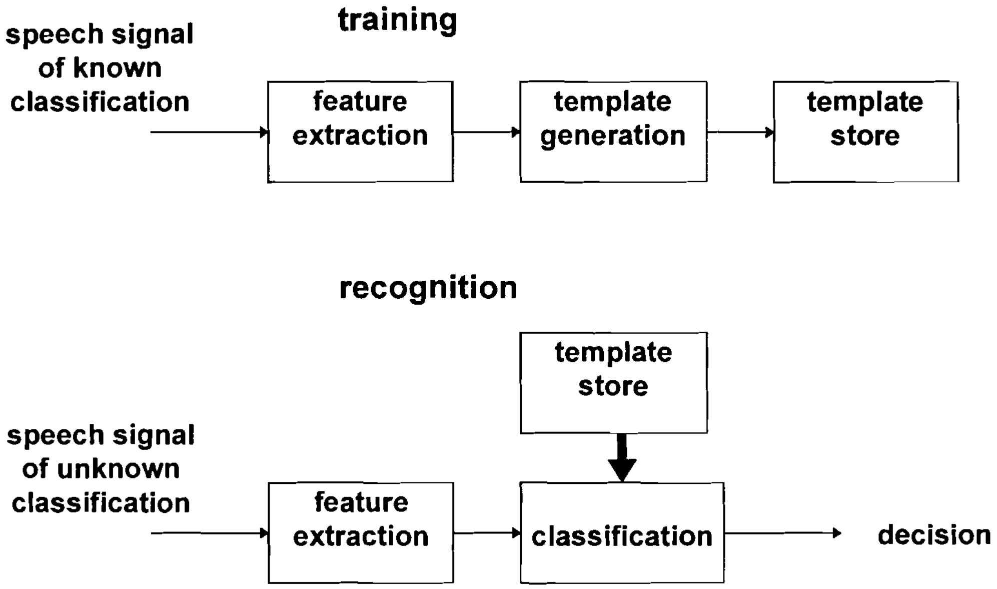


Figure 2-2, a pattern recognition system in training and recognition mode.

In most pattern recognition methods first some form of data reduction on the speech signal is performed. In that case the classification of the speech signal is based on the classification of some features describing the speech. These features are obtained by preprocessing the speech signal as described in paragraph 2.2: signal analysis and normalization.

In **training mode** several patterns corresponding to the same pattern class (i.e. the same word in the vocabulary) are used to create the reference template, representing the features for that class of patterns. The template can either be a model in which parameters are tuned by each train-pattern (see the discussion on hidden Markov models), or it can be an ordinary set of feature vectors (forming the pattern) derived from some kind of averaging technique.

In **recognition mode**, the classification of an unknown incoming speech signal is done by some means of similarity measurement between the features of the incoming speech signal and the stored reference features of each template. The decision logic eventually decides to which class of patterns the speech signal belongs. The decision is of course made using the similarity measurement to each reference template; next to this a certain threshold can be applied.

2.4.2 Template matching.

The most straightforward technique to perform speech recognition on the basis of pattern recognition is "template matching". This method simply stores the feature vectors during the training of the system, and matches the feature vectors from the unknown speech signal to each reference template.

In this approach the template pattern for each word, is composed of feature vectors. A feature vector describes the properties of a time frame of say 20 milliseconds. If the normalized time for a template is 1 second, then 50 feature vectors make up a complete template. In this way each template forms a matrix of 50 columns and a number of rows equal to the number of features extracted from the speech signal (see figure 2-3).

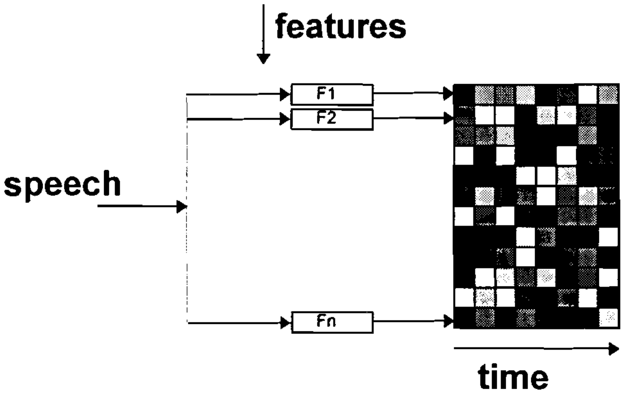


Figure 2-3, speech pattern based on feature vectors.

Upon recognition, the similarity between the speech signal and the several reference templates is expressed as a distant metric D(x,y) that measures the distance between speech pattern x and reference pattern y. There are a number of distance measures, e.g. the Euclidean distance, Mahalanobis distance, etc. [Rowden92]. The template pattern having the least distance to the speech pattern is said to be the recognized word (possibly the distance must exceed a threshold).

Next to the linear time and volume normalization performed in the preprocessing of the speech signal, a nonlinear time normalization may be needed to match the speech pattern on a template. Dynamic Time Warping (DTW) is a technique applied during template matching to compensate for interspeech rate variations. This technique allows parts of words to be stretched or normalized differently then other parts in the same word, i.e. non-linear time alignment [Neutelings87].

2.4.3 Statistical pattern recognition.

Next to the "simple" template matching technique, a more sophisticated method to perform pattern recognition based on statistical techniques is applied. "Speech differs from most signals dealt with by conventional pattern classifiers in that information is conveyed by the temporal order of speech sounds. A stochastic process provides a way of dealing with both this temporal structure and the variability within speech patterns representing the same perceived sounds." [Cox90]

The most widely used statistical method to perform speech recognition is the hidden Markov model (HMM) approach; extended information on HMM's can be found in [Cox90], [Moore92] and [Rabiner93]. In the HMM approach an utterance is seen as a sequence of articulatory outputs. A HMM (a stochastic automata) generates a sequence of observation vectors with a certain probability. The elements of an observation vector are the outputs from each state of the automata; a sequence of observation vectors can be seen as a representation of the sequence of articulatory outputs constituting the utterance.

The probability with which a specific sequence of observation vectors is generated, depends on the initial state of the automata, the state-transition probabilities and the output probabilities in each state of the automata. The idea now is to generate a HMM during training, for each specific word in the vocabulary; i.e. the different probabilities in the automata are tuned. Ergo, a reference template is a HMM representing one word of the vocabulary.

When recognizing a word, the likelihood of each HMM generating the sequence of feature vectors representing the speech signal is computed. The unknown utterance is classified to the class of words represented by the HMM that has the highest probability of generating a sequence of observation vectors equal to the sequence of feature vectors of the speech signal. The sequence of feature vectors was derived from the speech signal in the signal analysis preprocessor.

3. The user interface.

Speech recognition hardware providers all claim for their chips to have a very high recognition accuracy, and always try to achieve higher recognition rates by improving the hardware. Subsequently, or better, next to this, the recognition accuracy can be much improved by applying some ergonomics. Next to improving the recognition accuracy, a good user interface is essential to provide for a pleasant device that can actually improve the quality of life, especially in the case of severely disabled. This chapter tries to highlight some important issue's concerning the design of a proper auditive user interface when using a speaker dependent isolated word speech recognizer.

In [Jones89] some important design guidelines are given. Some particularly important guidelines for the application of an environmental control system for disabled users are:

1. a special command vocabulary should be designed for voice input,
2. provide feedback about the recognizer's activities,
3. provide representative template training.

In addition to these design guidelines given in [Jones89] we also should pay attention to the choice of type and position of the microphone. Especially physically disabled persons should not be bothered too much by an annoying microphone. This gives us a fourth design guideline:

4. Pay attention to the type and position of the microphone.

In the next paragraphs the design guidelines given above are discussed in detail, and the consequences for the design of the speech recognizer in the environmental control system are given.

3.1 Vocabulary design

The problem of identifying words correctly is mainly influenced by two factors: the number of words in the vocabulary and the "acoustic distance" between those words. First the issue of "acoustical distance" is wider explained, and next the impact of the vocabulary size is discussed.

3.1.1 The acoustic distance between words.

The first important issue in vocabulary design is the "acoustic distance" between the words in the active set of the vocabulary. Words having a large distance to each other do not sound alike at all, whereas words that have a small acoustic distance can easily be mixed up, as they sound almost the same. The military set: "Alfa, Bravo, Charlie,..." is a good example of a vocabulary, with words having large distance for human ears [Green83]. But, take care, not all words that have maximum separation for the human ear have similar distance for an electronic speech recognizer and vice versa! A good example of this common mistake is given in [Noyes89]: the users of the speech recognizer described in this article found it very frustrating that there was no connection between acoustical similar sounding words (for human ears) like "that" and "hat", and those words confused by the system. For example, if a user gave the command "fan", and the speech recognition system heard "alarm", then the alarm was continually activated, and the person which should help, might cease to respond to the alarm.

Therefore the choice of vocabulary words should not be left to the users; the designer should propose the words to be used after he has tested the words with the particular speech recognizer in his design (note that not all speech recognizers have the same optimal vocabulary). On the other hand the choice of command words should not be too obtrusive, because one cannot propose a vocabulary that will be best for all users, as some individuals will find some words easier to pronounce and remember.

Because of this, one often has to deviate from the familiar expression for a command in the design of the command word vocabulary, because of the desired distance between words. If one has to choose command words other than the normally used words for a particular command, then make sure that these substitutes fulfill the criteria of being **familiar**, **meaningful**, and **acoustical distinct** (for the speech recognizer used). For words to be acoustical distinct [Green83] gives some practical rules:

a. The beginning of a word should be loud enough, therefore it helps to choose words starting with a plosive consonant (p, t, k, b, d or g). Notice that in Dutch the "g" is not a plosive consonant.

b. For most recognizers, short words with (distinct) long vowels are recognized best.

c. Take words that start with a stressed syllable, so that the sound of the word is loud enough to pass the threshold of the speech recognizer. If one chooses words with the stress on the second syllable, sometimes the speech recognizer will hear the first syllable and other times it will only hear the second syllable and may recognize the wrong utterance.

d. Care should be taken when using words with plosive consonants in the middle of a word, these plosives create a gap within the word. The recognizer could see this gap as a pause between two distinct words. Another problem when using words with plosive consonants in the middle is that a person doesn't always speak out the two syllables in the same pace; so these kinds of words put a high burden on the recognizers job, and should be avoided. A sensible pause time between two explicit different command utterances is about 100 ms, but this depends on the settings of the speech recognizer.

3.1.2 The number of words in the vocabulary.

To get the most out of a speech recognizer one should keep the number of words in the vocabulary as small as possible; the chance that an utterance will be misrecognized increases with the size of the vocabulary. The number of words in the active vocabulary set can be kept as small as possible, by the use of a smart command syntax. In using this command syntax, we can take advantage of the fact that not all command words are functional in a certain state of the system, hence the number of words to be matched in the active vocabulary can be small. For example: if a user wants to close a certain door, he gives the command "door—close", obviously, once the command "door" is given, the next sensible command word can only be "open" or "close". So we don't have to make the recognizer check all the words in the vocabulary, it only has to check whether the command is "open", "close" or an erroneous command. Using this technique of a command tree, shown in figure 3-1, the (number of words in each) vocabulary set in each state of the system is largely reduced, and by that, the recognition accuracy is enhanced; see chapter 5 on testing.

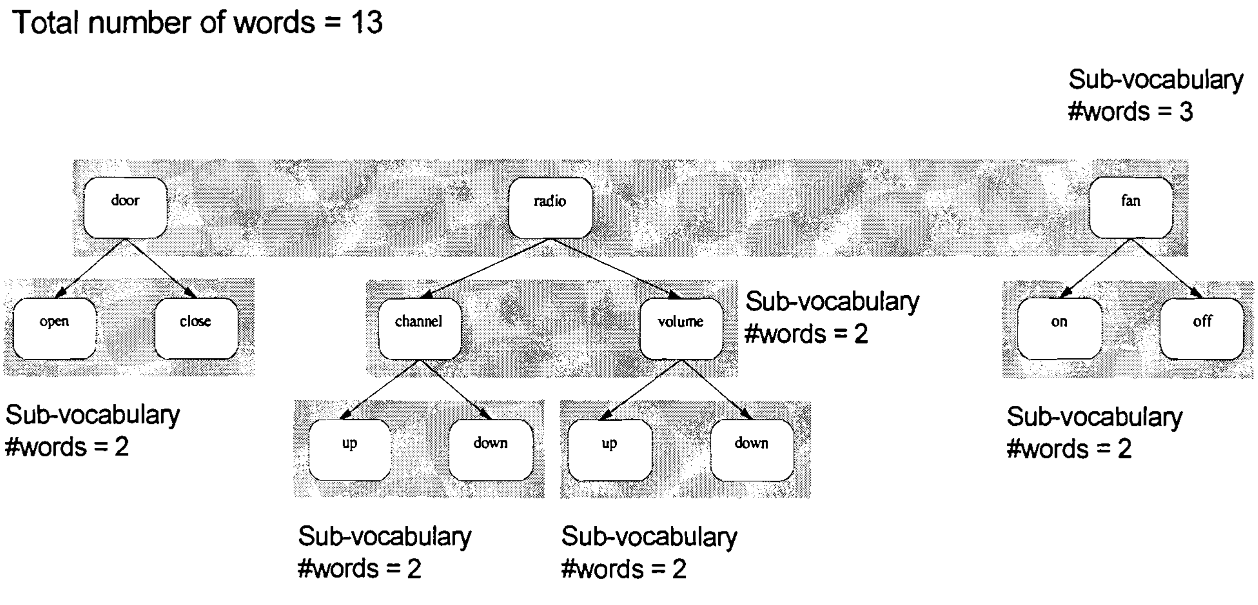


Figure 3-1, the smart use of a command tree decreases the number of words in the active vocabulary.

Another issue related to the choice of a command vocabulary, is the choice for a wake-up and a sleep command. The wake-up command is used to activate the speech recognizer to be sensitive for all normal command words, see figure 3-2.

The speech recognizer shouldn't always be sensitive for all the command words in the root of the command tree, as the number of false alarms increases rapidly with the use of more templates to be matched. When a speech recognizer has to check the utterances it captures against only one specific conspicuous template (for example a three digit code), then the threshold for this particular template can be low, because most utterances heard by the speech recognizer don't resemble this sole template at all, because of its conspicuousy. This conspicuous command word therefore can be used as a trigger to activate the speech recognizer and make it sensitive for all normal command words in the root of the command tree.

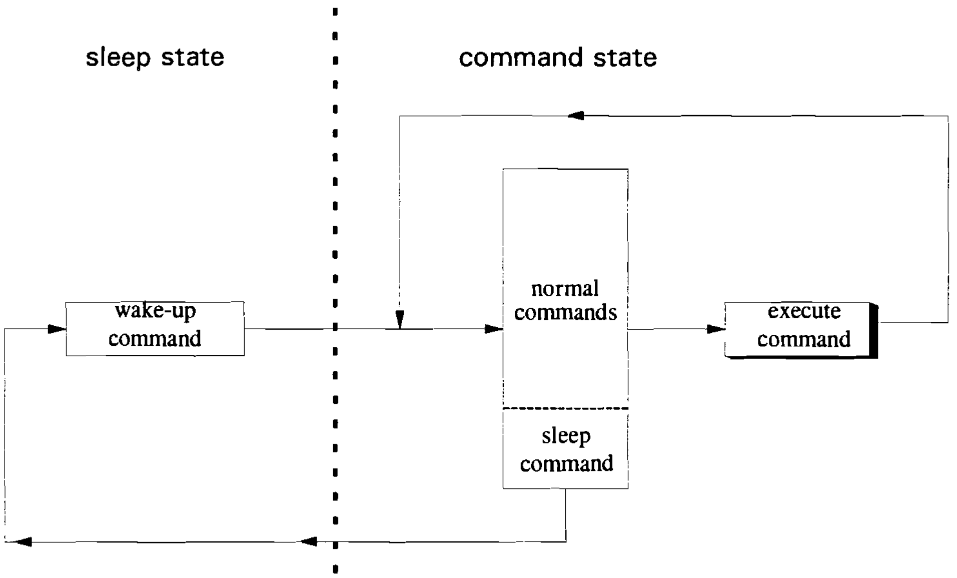


Figure 3-2, the use of a sleep state.

Another special command is needed to put the speech recognizer back into the sleep state, i.e. a "sleep command". One way of doing this, is to be silent for a certain period of time, another could be the use of an explicit sleep command. The latter is preferable, because there is always a certain amount of background noise in the speech recognizer's environment, and the user has to be silent, and wait for the time to be passed. Unlike the advantage of just having to check for one specific word in the wake up case, the sleep command must be accepted in the command state (see figure 3-2) amongst all other normal commands. The effect of an attention word (like the wake-up and the sleep command) can even be enlarged by pronouncing this word twice (see the use of the wake-up command in chapter six and seven).

3.2 Provide feedback.

For an auditive user-interface (i.e. the interaction is based on audible commands, and their responses), it is very important to provide for immediate feedback, because sounds are volatile, and the user can't "look" back (i.e. inspect) at the things he just said. It is important that the user should be supported as much as possible in keeping track of the systems state.

There are three issues involved in giving appropriate feedback:

⇒ **What** This concerns the issue of what kind of feedback should be given.

⇒ **When** Another important topic concerns the timing of the feedback given.

⇒ **How** The third important point is the modality of feedback; visual, audible, ...

3.2.1 What should be fed back.

The minimal kind of feedback the user should get from the recognizer is an indication when the speech recognizer is ready to listen, and a signal whether an utterance is recognized or not. This can be achieved in two extreme varieties:

I. The most straightforward kind of feedback is to provide for non at all, except for the action followed by the recognized command word, i.e. implicit feedback. For example, the user says: "light"-"on", subsequently the speech recognition system opens the door; by this action the user notices that the wrong command was recognized. If we want a more graceful kind of feedback, then explicit feedback should be used.

II. The other extreme would be a means of feedback that directs the user to speak out an utterance in such a way, that the recognizer could better recognize the spoken word. This would be some means to show the user in what way a spoken utterance resembles the stored template, so that the user immediately notices in which way he could better pronounce the word, and thus improve the recognition process.

A reasonable compromise between these two extremes would be a reject indicator, signaling that no command is recognized, an echo off the utterance recognized when a command is recognized, and a ready indicator activated when the recognizer is ready to listen to a command.

3.2.2 When should the feedback be given.

With the use of a command tree, as discussed in paragraph 3.1.2 a bit more sophisticated feedback mechanisms are possible. One could think of a feedback signal to indicate that the first word of a command was recognized correctly, and that the system is waiting for the next command word (the next branch in the command tree) to be uttered. In this case the feedback is given concurrently and it can be used for regulating the timing of input. But this brings up another problem: the speech recognizer needs more time for the recognition of an utterance in a large vocabulary than for a command word in a small vocabulary, as it has to match more different templates. So the time before the feedback is given, upon recognition, is different for each specific vocabulary. This means that the timing of the feedback is unpredictable, and therefore annoying to the user. One way to overcome this problem is to make the delay in all cases as long as the longest delay. The fact that this slows down the user is of no big importance to handicapped users, especially when using a command tree with few levels.

3.2.3 How should the feedback be given.

Next to the problem **what** type of feedback should be used and **when** the feedback should be given, is the issue **how** this feedback is given to the user. The choice of modality in which explicit feedback is given, is restrained to visual, auditory or a combination of both. In the application of voice controlled domestic appliances by disabled, this choice is obviously restricted by the kind of disability of the users (you can't use visual feedback if the user is visually disabled). Another issue that determines how the feedback is given, is the portability of the recognition device, a person can't lug a 14 inch monitor around the house, whereas auditory feedback is very well suited to be used in a situation where the user wants to move through the house. If the user is mobile, either the feedback has to be auditory, or the person has to use a portable (small) display, since he would generally be outside the range of a fixed display. The major disadvantage of auditory feedback however is that it tends to disrupt any other auditory information that is kept in the memory of the user [Jones89]. But with a small command tree the user doesn't have to remember a range of command words following each other.

When using a large command tree as described in the previous paragraph care should be taken that the short term memory of the user isn't overburdened. An individual can keep approximately 5-8 items in his short term memory[Bradford95], so if one uses a command tree with more than say 4 levels in hierarchy, an extended means of feedback should be provided, so that the user can keep track of the state of the system. When using a small command tree with few levels a simple indication whether the command is correctly understood (until now) should be sufficient. Another solution is to provide for both a small display, and the mentioned auditory signals [McCauley84], in an optimally performing system, the user would almost never hear a tone, and there is no need to inspect the display. In the case where the recognition accuracy is low (new user), full feedback information is necessary (display).

3.2.4 Error handling.

Another important topic in the feedback issue is to provide simple ways to get out of an erroneous situation.

Users may find themselves in blind alleys not knowing how to get out of an interactive process, or they may say one thing, and mean something else. To deal with this kind of situations, in general, one should provide for a means to allow the user to back up one step and provide a means for the user to get back to the beginning. In the case of a small vocabulary, and a shallow command tree the first option may be omitted. If a means to back up one step is provided, one also has to provide for a constant indication at which the user can see in what state the system is, otherwise the back up option makes no sense. For a small command tree, the overhead used to provide for this option may be too big. In that case a user could better simply go back to the begin (the root of the command tree), than have to remember how he could back up one step, what the next command should be, and how to perform these actions. More specific, the back up (to the beginning) option in a small command tree, can be implemented as follows:

One could think of making use of a minimal delay time that is restricted, between two command words that follow each other in a command. For example, if a user wants to turn on the radio, he should give the command: “radio”-”on”. Now say the user mistakenly says: “television”, in stead of radio; if he notices his mistake in substituting the two devices, he can be silent and wait for the minimal delay time to be passed. Meanwhile the speech recognizer has given a signal (auditory or visual) to indicate the user that the first command word was understood properly and that it is waiting for the following command (the next branch in the command tree). But if the user doesn’t reply with this next command, the speech recognizer knows that the user made a mistake, and the speech recognizer’s state is set back to the root of the command tree. An even better solution than the passive silence, would be an active command (like “escape”) from the user to go back to the root of the command tree. In paragraph 6.2 the use of such “cancel” command is applied.

3.3 Template training

The purpose of template training is to store relevant utterances of the command words to be recognized. In speaker dependent word recognition this template training is obviously done by the user himself, whereas in speaker independent voice recognition the templates are trained by a large number of persons and prestored in the factory, see paragraph 2.1. In this paragraph only the issues concerning speaker dependent voice recognition are considered.

The most difficult part of template training, is to store templates that resemble maximal the command words, spoken out in operational usage of the system. Traditionally the template training is done by reading all the command words from a list, a few times one after each other in a laboratory environment. Subsequently, the average of these utterances of the same word are stored as a template. This "list reading" leaves a lot to be desired, because the uttered command words don't resemble the commands spoken out in operational use at all. This is because of several reasons [Jones89]:

- The recorded speech is sensitive to the phonetic environment in which it is spoken. The templates should be recorded in the operational environment, with all it's acoustic features, and background noise.
- Speech is sensitive to the semantic context, since the way speech is uttered reflects the meaning that is being conveyed. Words sound different, being read from a list (dull), than being used as a command in a real life situation (more expressive).
- The state of the individual is different in a laboratory environment, where he just has to read out a list of commands, than in a real life situation, where he might be more relaxed, or maybe more stressed. Another point is the speed at which the commands are spoken.
- The method of averaging a few utterances of the same words doesn't always result in a good template. Green et al in [Green83] found out that the usage of storing more non-averaged templates of the same words led to significantly better recognition rates.

But what can be done to overcome these problems? Well, that is obvious, the training session should be made more life-like, i.e. the training situation should resemble the real life application as much as possible. This can be done as follows.

The semantic problem can be tackled by not having the user read out the command from a list, but by having him perform a virtual task, very much alike the real task. In this case, where the speech recognizer is used for controlling domestic appliances, a virtual house could be created on a PC. The user finds himself in this house, and has to perform some specific tasks requested by the system.

For example, the user is in the kitchen of the house, and is told to go to the living room to watch TV. Just like in a real life situation, the user has to think of the order of subtasks he has to perform, to achieve the goal. In this way the user is distracted from the fact that he is training the speech recognizer, i.e. hidden training. The user sees himself in the kitchen, and knows he has to open the kitchen door to go to the living room. Therefore he speaks out the command: "kitchen door—open" in the same way he would have said it in operational use of the system.

By using virtual tasks like this, the issue of semantic and the stress factor difference between the training session and the operational situation can be dealt with. One aspect of this kind of training however, should not be overlooked: the commands the user gives in his virtual task must be 100% predictable, otherwise the wrong utterance is stored as a template for another word. But with a smart design of such subtasks and a serious user, this should be possible.

The hidden training described above should only be done at the first time the speech recognizer is trained. Nevertheless, retraining is needed, as the human voice tends to change over a period of time, and the acoustical features of the operational environment can change. To have the recognizer trained in the same way as the first time may be to much of a burden to the user, therefore an adaptive method to update the templates may be needed. A sophisticated method of continuous retraining by means of an adaptive user interface is given in [Green83]. In this method multiple templates of the same command word are updated adaptively, to compensate for the smooth drift in the voice of the user.

3.4 Microphone placement

The choice, what microphone has to be used, and where to put it, is an important factor determining the quality of the speech signal that is put into the speech recognizer. The most important microphone issue is that the signal from the microphone is consistent; i.e. the type of microphone and it's position, should resemble maximally, in training and in operational mode. Next to this important topic, the type and position of the microphone are determined by some specific user dependent properties.

In the application of domestical control for disabled persons, roughly, two kinds of situations can be distinguished.

1. An **immobile user**: the disabled user is always at a fixed position in his room (e.g. in bed).
2. A **mobile user**: the user is able to move through his house (e.g. in a wheelchair).

1. An immobile user.

If the user is in only one room and most of the time at the same place, (this is a situation not unfamiliar in the domain of the severely handicapped), it is easy to place the microphone at such a place that the distance and the angle to the user is always the same. The microphone can either be **attached to the user**, or be located at a **fixed position in the room**.

If the microphone is at a **fixed position in the room**, it could either be mounted nearby the user; in this case it might be impeding to his freedom of action, or it can be mounted at a distance from the user. The latter has the disadvantage of a longer acoustical channel, and as a consequence, the signal may be distorted by the acoustical transmission characteristics of the room.

In the situation where the microphone is **attached to the user**, we don't have to worry that much about the varying acoustical characteristics of the room, for the distance between the microphone and the users mouth is small, and fixed. The use of a headset should be avoided, as it is very obtrusive to the disabled user; a tie clip microphone somewhere attached on the users breast, is more suitable. The tie clip microphone doesn't bother the user in his freedom of action, and the signal lead can easily be disguised under the user's clothes. The position of the microphone should be maintained consistently in the use of the system, as variations in microphone placement result in differences in the acoustic input to the system.

2. A mobile user.

In the situation where the user is able to move through the house (in a wheelchair), again there are two options: the microphone can either be at a **fixed position** in the room, or it could be **attached to the user**. For a user who is mobile, we have to worry even more for the microphone not being too obtrusive.

If the microphone is at a **fixed position** in the room, and the user is mobile, the distance between the user and the microphone constantly changes; the same holds for the angle between the user and the microphone. This puts a high burden on the speech recognizer, as the characteristics of the acoustical channel are different all the time. This problem could be dealt with, training the speech recognition system from different places in the room, nevertheless, it is obvious that a microphone at a fixed position to the (moving) user is far better.

A tie clip microphone, **attached to the user**, seems to be most suitable for a moving person, as it doesn't bother the user too much, and it is always at a fixed position to his/her mouth.

A microphone attached to a moving user rises another problem; how should the signal received by the microphone be conveyed to the speech recognizer ?

Position of the speech recognizer.

In the situation where the user is always at a fixed position in the room (e.g. in bed) the speech recognizer can be placed nearby this place, and there is no need to worry about its physical dimensions, the power it drains, and the way the system communicates with the environmental control system. If the user is mobile, and carries the microphone somewhere attached to him, we have two choices for the location of the speech recognizer:

1. it can be positioned near the user, i.e. he has to carry the speech recognizer along,
2. the other option is, to place it somewhere at a fixed place in the room.

In both cases there is some kind of communication necessary; between the speech recognizer and the environmental control system in the first option, and between the microphone and the speech recognizer in the second case. Both put a burden on the usage of energy (batteries), still, this problem can be overcome by making use of the existing batteries on a wheelchair.

1. The speech recognizer is carried along.

When carrying the complete speech recognition system, there has to be some send equipment to transmit the recognized commands to the environmental control system. There is another disadvantage of carrying the complete speech recognition system, over carrying just the microphone: the complete speech recognition system has greater mass and volume; this could be annoying to the user. An advantage however, is that the communication between the speech recognizer and the environmental control unit can be simple and robust; for the amount of data to be passed is small, and can easily be coded in a robust way (see chapter seven on implementation details).

2. The speech recognizer is at a fixed position in the room.

This means that the complete electrical signal from the microphone has to be converted (modulated) and transmitted on a r.f. (radio frequency) carrier. The fixed speech recognition system would then be coupled to a r.f. receiver, to obtain the signal from the microphone. Off course, the r.f. signal can be distorted; ergo, the speech signal will also be distorted. The distortion of the transmitted speech signal wouldn't be that unpleasant, if the distortion were the same, each time the command was given; including the moment the speech recognizer was trained. However, because of all kinds of varying conditions in the speakers environment (the position of the speaker, the presence of electrical magnetical polluting machines, etc.) the received r.f. signal from two equivalent utterances can be different, and thus the recognition accuracy will decrease!

4. Selection of the speech recognition system.

An important part of a voice controlled environmental control system obvious is the sub-system that performs the speech recognition. The need for the complete system to be small, low cost and stand-alone, restrains the speech recognition subsystem to the same criteria. Therefore, the selection of the speech recognition system is an important issue.

A great restriction is the fact that the speech recognition system should not be implemented on some kind of computer system. This criterion excluded many commercially available sound cards, and speech recognition boards for utilization in a personal computer or in other computer systems.

4.1 Requirements for the speech recognition system.

Being at the very beginning of the orientation on speech recognition systems, and not wanting to restrict the choice too much yet, the specifications of the speech recognition system are only loosely formulated. Apart from being stand alone, there are some strict criteria the speech recognition system has to meet:

I. the speech recognizer should provide speaker dependent isolated word recognition for a vocabulary of at least 30 words,

II. the system should not drain too much power,

III. it has to be low cost (< $100).

With nowadays speech recognition technology, the first criterion is not very hard. But together with the demands of being small, stand alone and low cost, the choice is not that simple.

In chapter 2 on automatic speech recognition, the various steps in performing speech recognition were discussed. The first part required for the speech recognition system is an electronic filter and gain unit, to process the microphone signal. Another indispensable part needed to convert the analog speech signal to its digital equivalent is the analog-digital (AD) converter. The speech analysis part, the normalization of the speech features and the actual recognition/classification part require many computational operations, data scaling and addressing, and complex decision making. At least some kind of micro processor is needed to perform these various computations.

For the system to be convenient in speech recognition, the processing of an utterance should not take too long; so the microprocessor must be fast enough to handle all the signal analysis and recognition computations. A digital signal processor (DSP) is a microprocessor that is dedicated to perform fast signal processing on large amounts of data. Apart from this advantage over conventional microprocessors, it is equipped with extensive peripherals (AD/DA converters, serial ports, parallel ports, etc.) to communicate with the outer world.

Due to all these advantages over conventional microprocessors, the choice for a DSP-based solution is easily made. The use of an application specific DSP, to perform speech recognition, places minimal burden on the host and other system components. That is: an application specific DSP can implement all speech recognition, while the high level operations to support sophisticated user interface and vocabulary management can easily be done by a simple microcontroller. At this point we have restricted the selection of the speech recognition system, to the selection of the application specific (i.e. speech recognition) DSP.

Prior to the design of the complete system, we are faced with evaluating and choosing a (speech recognition) DSP, that offers speaker-dependent isolated word recognition. Especially the criteria of cost, recognition performance, robustness, and power consumption should be considered well.

4.2 Selection of the speech recognition processor.

Next to the speech recognition processor a speech recognition system using a dedicated DSP is usually build up with an input stage to perform filtering and (automatic) gain control, a micro controller, memory and a crystal (clock). Considering the memory; RAM is required to hold the captured utterances, and to store intermediate arithmetic results, and ROM is applied to hold the program-code for the speech recognition algorithm.

As IC manufacturers tend to advertise their products more and more on the internet, most of the information on today's speech recognition was found by investigating the World Wide Web (WWW), and a newsgroup dedicated to speech technology[1]. Most helpful in the selection of the proper speech recognition IC was a data-base[2] of commercially available speech IC's made by J.P. Lereboullet.

In the next few paragraphs a short description of the most important (to the project) speech recognition chips available today is given.

[1] URL: usenet://comp.speech

[2] URL: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/info/VoiceRecognitionProcessors

4.2.1 Five candidates.

The investigation for speech recognition chip's, led to five important candidates:

DVC306 from DSP Communications Inc.,
D6106 from DSP Communications Inc.,
HM2007 from Hualon Micro-electronics,
MSM6679 from Oki Semiconductors,
RSC-164 from Sensory Circuits Inc.

All these IC's were introduced in 1995 or 1996, so they are definitely state-of-the-art. They all comprise the possibility of speaker dependent, isolated word speech recognition, and don't need a computer in operational mode. Before we take a look at the highlights of the individual chip's, the following abbreviations are explained:

SD:	speaker dependent,
SI:	speaker independent,
IWR:	isolated word recognition,
μC:	micro-controller,
CODEC:	coder decoder: this circuit converts an analog speech signal to a digital representation of this speech signal,
XTAL:	crystal.

In appendix 1 a detailed overview is given of the features of the five speech recognition chip's.

The **D6106** from DSP Communications Inc. is a SD IWR speech recognition chip, introduced in 1995. This chip is optimized for noisy environments, and can keep up to 128 words. Fore this, it needs maximal 2*128 kB SRAM, and 2*128 kB (EP)ROM. It has an 8-bit bi-directional host-interface, and an 8-bit external memory bus. To built a complete speech recognition system, one has to add: a PCM CODEC, an 8-bit μC, some (EP)ROM and SRAM, a 29,5 MHz XTAL and a microphone.

The successor of the D6106 is the **DVC306** (also from DSP Communications Inc.), it offers both SD and SI voice recognition, and continuous speech recognition by means of key-word spotting. It provides for a way of prompting and verification by means of speech synthesis. Just like the D6106 it is said to be extremely robust for noisy environments and in addition, it can be used for memo pad recording of small messages. The DVC306 can comprise up to 1 MB of external SRAM to store 16-128 words for up to 8 users in SD mode. This speech chip needs a CODEC, an μC, a XTAL and external memory (both ROM and SRAM) to operate as a speech recognition system.

The voice recognition chip from Hualon Micro-electronics, **HM2007** offers SD IWR for 40 words of 0.9 sec or 20 words of 1.9 sec. To achieve this, the chip needs an external 8 kB SRAM chip and a standard 8-bit μC. An electret microphone can be directly connected to the chip and there is no need for an additional CODEC. The 8-bits bi-directional databus can be used to give the result code, which indicates the confidence level, to the host μC, or the bus is used to up- and down-load voice data.

Oki Semiconductor's **MSM6679** is a "ready-to-go" voice recognition processor offering SD and SI word recognition. It is standard preprogrammed with 3 SI-vocabularies containing 20 templates for digits, "Yes", "No", "dial", and other custom telephone commands. The external memory can be extended to 64 kB SRAM and (EP)ROM per bank, which offers the possibility of storing 61 SD words per vocabulary. The MSM6679 provides user selectable feedback, either a voice prompt or a simple beep, to verify that a voice request has been understood. The chip offers both a serial and a parallel host interface to a host μC or a computer, and has an on-chip interface to Oki's speech synthesis IC's. This chip makes use of ADPCM coding, and provides for a voice output of this kind.

Unlike the former chips, the **RSC164** from Sensory Circuits Inc. is not based on a digital signal processor but it makes use of a neural network. The chip incorporates an 8 bit, 4 MIPS μC, circuitry for automatic gain control, analog to digital conversion, 384 bytes of RAM, 64 kB ROM and 16 programmable input/output lines. The on-chip RAM can hold up to 40 seconds of audio. The chip is SD, 1-10 words per vocabulary and SI, 2-20 words per vocabulary, where the number of vocabulary's is only limited by memory size. The RSC164 also offers speech synthesis, 4 channel music synthesis and a record-playback facility using data rates of under 14,000 bits per second.

4.2.2 The final speech recognition chip.

All the five IC's proposed meet the minimal features required; so in this last stage of the selection procedure the selection has to be made on some other bases. The recognition performance of course is an important issue in selecting the speech IC.

At this point we have come to a big problem, because there is no standard in defining the "recognition rate", or "accuracy" for a speech recognition system. Each manufacturer has its own definition of "accuracy" or "recognition rate", and all manufacturers present this sort of meaningless numbers for their products. They all claim to have a "recognition rate" of over 95%, but none of them discusses the way these marks of quality are obtained. In chapter five on testing, this topic is further discussed. But by now, the choice has to be based on some other aspects.

Another important issue obviously is the availability of the chip at a distributor nearby, to give technical support. Being at the very start of this research project, decisions that are made now, have an effect for the next few years. The choice of the speech recognition processor on which the system will be based, is such an important decision. This means that the speech IC finally chosen should be supported in the future by a reliable manufacturer (or distributor). As most manufacturers are located in the USA or in the far east, we have to look for distributors in (at least) Europe.

After an extensive search for information on availability of distributors for the five speech recognition IC's, it came out that only Oki Semiconductors has many reliable distributors in our area (the Netherlands). All the other manufacturers could not offer support in Europe, or even didn't answer when asking for a distributor list. At this point the decision was made for the MSM6679 Voice Recognition Processor from Oki Semiconductors (MSM6679). In the next paragraph the features of this chip, and the consequences for this project are further discussed.

4.3 MSM6679 voice recognition processor

In this chapter the features of the MSM6679 Voice Recognition Processor from Oki Semiconductors (further referred to as MSM6679) are overviewed. First a functional description of the MSM6679 is given, in which the most important features of this voice recognition processor are shown. Next to this, the way the processor operates in slave mode is further described.

At the end of this chapter the consequences of using this particular voice IC for the project are investigated. Most of the information given in this chapter is retrieved from the MSM6679 datasheet [MSM6679-96], and the EVA KT 6679-2 evaluation board datasheet [MSM6679-EVA-96], more specific information on the MSM6679 can be found in these documents.

4.3.1 Functional description of the MSM6679.

The MSM6679 is a host driven speech recognition processor; that is, it is commanded by a host system to perform speech recognition (see next paragraph).

The MSM6679 can perform five functions:

- speaker-independent isolated word recognition (SI-IWR),
- speaker-dependent isolated word recognition (SD-IWR),
- solid-state sound recording,
- sound playback,
- speech synthesis.

Isolated word recognition

The MSM6679 can perform both Speaker-Independent (SI) and Speaker-Dependent (SD) speech recognition. SI vocabularies are factory pre-trained and stored in the MSM6679. For SD recognition, each recognized word must be enrolled (pattern recognition approach) in the MSM6679 vocabulary. During training a composite template is created from multiple recordings of the same word, which is then stored in SRAM or FLASH (EEPROM) memory, and can be updated any time. The MSM6679 performs speech recognition using the statistical pattern recognition approach; see chapter 2 on automatic speech recognition.

During both SI and SD speech recognition, the following steps are performed:

1. After external band-pass filtering and automatic gain control, the MSM6679 converts the analog speech signal to (digital) PCM (pulse code modulated) samples.
2. The MSM6679 extracts significant features from the sample data by frequency and time-domain analysis; that is signal analysis. The algorithm is based on dynamic time warping (DTW) and hidden Markov models (HMM), and patented by Voice Control Systems Inc. (VCS).
3. The MSM6679 compares the analyzed input with each stored reference pattern (template), weighing the significance of similarities. A score (expressed as distance) is generated for each template in the actual vocabulary.
4. The vocabulary word that achieves the highest score (or smallest distance) is judged to match the input phrase, assuming that the score exceeds a pre-determined threshold.
5. Via a special host command, the MSM6679 can return the scores of the input against all defined vocabulary templates for SI and SD speech recognition. This allows external host software to select the next best match, if (for example) the closest match is not contextually logical.

SI recognition

The MSM6679 is standard supplied with a pre-defined SI vocabulary. Each template in this vocabulary is trained with a wide variety of speakers, so the stored reference templates will closely match most utterances of the words spoken by anybody. Oki can implement a custom SI vocabulary for a fairly high price, but the standard embedded vocabulary can be used for many (English language) applications. The standard embedded SI vocabulary contains 20 words; among which the digits 0-9, “Yes”, “No”, “dial”, “directory”, and more words used in telephone applications.

SD recognition

In SD recognition mode, the MSM6679 can be trained to recognize up to 61 words. The MSM6679 can support multiple speakers by switching vocabularies, but only one speaker's vocabulary can be active at a time.

The user enrolls a word in the MSM6679's vocabulary initially, by recording the word three times or more. In addition to initial enrollment training, adaptive template updating can further improve the recognition accuracy. The host application could update templates by first asking the speaker to confirm a recognized word with a "yes" or "no".

Solid state recording/playback

As well as providing speech recognition capabilities, the MSM6679 contains a solid-state recorder/player. To facilitate feedback in SD recognition, the chip supports recording and playback of “name tags”. Name tags are used to confirm correct recognition, or to signal the user that the speech recognizer has captured the wrong utterance. The MSM6679 stores the name tags using an adaptive differential pulse code modulation (ADPCM) compression algorithm. This is done at rates of 28 kbits per second. The various record and playback operations can be performed using the host commands.

Speech synthesis

For speech-synthesis requirements, the MSM6679 also provides a MSM665x (Oki speech syntheses chip) control interface for external speech synthesis. Depending on the type of speech-synthesis chip used, up to 1 hour of speech (sampled at 4 kHz) can be synthesized! Next to this, the MSM6679 contains a tone-generator, to produce standard beeps and DTMF dial tones.

4.3.2 MSM6679 slave-mode API

The MSM6679 is a slave mode speech recognition processor, so it can be commanded by a host system to perform the various functions discussed in the latter paragraph, or to upload or download speech data. This paragraph briefly describes the slave-mode application protocol interface (API). The interface manages the communication between a host application, and the MSM6679. The host application can either be a computer system, or some kind of micro-controller system (μC).

The commands used to instruct the MSM6679 consist of an opcode and an operand. The opcodes for the MSM6679 consist of exactly four bytes with values between F000 (hex) and FEFE (hex), and the operands can be of variable length. According to the given command, the return code from the MSM6679 generally consists of the same opcode, followed by data indicating success or failure of the operation. A detailed description of all commands and the possible return values from the MSM6679 can be found in [MSM6679-96]. A mnemonic language (and parser) called “Spraakwater” was designed to program the MSM6679 with commands (opcodes and operands) that are more meaningful than the bare four digit opcodes. A detailed description on the syntax and semantics of “Spraakwater” is given in [Custers96]

The communication between the MSM6679 and the host system can be done through the parallel port, or the serial port of the MSM6679. This offers a flexible way of communication between the host system and the speech recognition processor.

4.3.3 Consequences for the project.

The selection of the MSM6679 speech recognition processor as the heart of the speech recognition unit of course has some consequences for the rest of the project

The most important issue is the necessity of a μC to command the MSM6679 and to perform the high level operations to support user interface and vocabulary management. Next to the μC, the MSM6679 requires an external memory and an electronic input stage (filtering, gain control). In paragraph 6.1, the structure of the complete system is further discussed.

The MSM6679 speech recognition processor offers far more functionality than the application of command driven environmental control requires. The extended features however are a most welcome extension to our application in this way:

- the recording and playback of name tags can be used to give sophisticated feedback,
- SI recognition could be used to initially enroll the voice of a new user by instructing the MSM6679 with SI voice commands. SI recognition might also be necessary to update templates, requiring "yes" and "no" from the user.
- the upload and download commands of the extended API can be used to retrieve speech data that might be convenient in performance tests.

Some tests to get a notion of the recognition performance for the use in environmental control must first be made, as we can not rely on the performance measures given by Oki (see the next chapter on testing). To perform these tests, an evaluation board [MSM6679-EVA-96] that can be hooked up to a PC is very useful. The evaluation board also allows us to get acquainted with the hardware of the MSM6679 and the control software to command the processor.

Regarding the costs; the MSM6679 is priced $25,- in large quantities (>10000), and the costs of the peripherals for the speech recognition system (μC, memory, etc.) will certainly not exceed $100,- .

The MSM6679 speech recognition processor will be available and supported as from February 1997 by many distributors in Europe (note that the European headquarters of Oki Semiconductor are located in Germany, Neuss).

5. Testing the speech recogniser

After the selection of the speech recognition sub-system (MSM6679 speech recognition processor), based on other criteria then speech recognition performance, some indication whether the recogniser is "good enough" should be obtained. In this chapter first the reason why the device is tested, and what has to be tested is discussed. Something is said about the test conditions, the test results are given, and the conclusions for the project are drawn.

5.1 Why testing?

At this stage of the project we have made a decision for the speech recognition system, based on availability and technical support of the chip at a reliable manufacturer or distributor nearby. Before continuing, some small tests to get an indication of the speech recognition performance of the selected chip should be carried out. This is the most important reason the tests were done. Next to this, the tests will be used to get an indication whether the words in the proposed vocabulary have large enough acoustic distance to eachother, and to notice what words are easily mixed up (read paragraph 3.1.1 on acoustic distance between words). Another aspect that can be learned from the test results is the influence of background noise, and the MSM6679's sensitivity to out of vocabulary sounds (i.e. sounds or spoken words that are not in the vocabulary; and thus should be rejected). Finally the tests give us a better understanding of the distance measures used in the MSM6679.

5.1.1 Recognition performance measures

All manufaturers claim to have "recognition rates" of over 95% for their speech recognition devices; so does Oki Semiconductors. It is very important to be sceptic about these recognition rates, and to bare in mind that they are retrieved conducting a very specific test. These tests are often conducted using highly trained speakers, which speak in a very consistent way. In 1969 R.S. Hyde stated a law (Hyde's law) covering this speech recognition performance issue as follows [Johnston96]:

"the accuracy of speech recognisers is 98%" ,

quickly followed by its first consequence:

"because speech recognisers have an accuracy of 98%, tests must be arranged to prove it."

It is evident that the "accuracy" or "recognition rates" published by manufacturers should be treated with Hyde's Law in mind. As a result, we may not fully rely on the recognition rates given by Oki (97%) in [MSM6679-96], but have to conduct a test that is more realistic on the recognition issues involved in this project.

As said before; the main reason for conducting the tests is to get an indication (no more, and no less) whether the MSM6679 is "good enough" for usage in a command driven environmental control system for disabled. The problem however is that we don't know what is "good enough" at this point. It is obvious that only the real level of user satisfaction is a measure for the project to be succesfull. The advantages of using the system, must be greater than the disadvantages (frustrations upon misrecognition) encountered.

It is important that the whole transaction of controlling domestical appliances, from initiation to the desired outcome is comfortable and free of frustration from the users point of view. This does not require 100% recognition accuracy, but well designed dialogues, and support software [Noyes89]

The speech recognition system is thought to be satisfactory for users if just one out of ten command words is misrecognised[1]. For example: the user says "door", and the recogniser understands "light" (this is a substitution error), or the recogniser doesn't "hear" a word at all (this is a deletion error). In [Johnston96] some conventions on recognition performance are given; in this paper Johnston states that "%correct is the figure which the user of an application experiences". The %correct performance measure is adopted here to get an indication of the user satisfaction[2]:

$$\textbf{\%correct} = \textbf{100\%} - \mathbf{E_{sub}} - \mathbf{E_{del}} \,,$$

where:

E_{sub} (a **substitution error**) is an error where a valid word is said, but the speech recogniser classifies it as an other valid word of the vocabulary,

and: E_{del} (a **deletion error**) is made when a valid word is presented, but rejected by the speech recogniser as being a word of the vocabulary.

5.1.2 What tests should be conducted

The previous paragraph described a measure to get an indication about the user satisfaction (i.e. %correct). Since we are interested most in that issue, we have to conduct a test to get the numbers for E_{sub} and E_{del}. A speaker-dependent isolated word recognition performance (SD-IWR) test, can provide the wanted numbers.

[1] This measure is both intuitively obtained, and determined by interviewing some people.

[2] In a more extended test, numbers on false acceptance, true acceptance, false rejection and true rejection should be incorporated [Johnston96].

To get a notion on the influence of background noise during recognition, one SD-IWR test should be done in an environment very much alike a studio; that is, minimal background noise, consistent position of the microphone compared to the mouth of the speaker, and a constant speech level. Another SD-IWR test should be done in a noisy environment, with various levels of speech.

Another distinction in SD-IWR tests will be made in the way the commands are spoken. One test will be performed using utterances that are read from a list of words (called listreading); the other test will have a speaker who gets a task, and speaks the (command) words in a more lifelike fashion (called lifelike speaking). The results of this test may give an indication on the issues of lifelike training in paragraph 3.4.

The acoustic distance between words can be acquired by a SD-IWR distance test. In this test, an utterance of each word of the vocabulary is offered to be recognized. In this way, the mutual distances between all words of the vocabulary are retrieved. The outcomes of this test should give an indication whether the words in the vocabulary are easily mixed up (small acoustic distance), or don't sound alike (for the MSM6679) at all (large acoustic distance). This test also makes us acquainted with the distance measures that are used in the MSM6679.

An “out of vocabulary” test may be conducted to get an indication on the sensitivity of the MSM6679 (and of course the used vocabulary) to “obtrusive” background sounds (e.g. sneeze, door bell, door slam, etc.) that might be recognized as being a valid command.

All specific tests described above, will be achieved by four general tests:

1. SWL; in this test, the speaker reads the words from a list in a “studio” environment,
2. SZL; here, the words are spoken more lifelike (using a task); the test is still performed in a studio environment,
3. KZL; the words are spoken using a task, but the test environment is noisy,
4. OOV; words trained in a studio environment are matched against “out of vocabulary “(OOV) sounds.

The MSM6679 produces distance scores [MSM6679-96] to give an indication of the “goodness of fit”. After each recognition, the distance scores for all words in the vocabulary can be retrieved from the MSM6679. These distances where used in all tests:

- first to get acquainted with the distance scores of the MSM6679,
- to obtain numbers on E_{sub} and E_{del},
- to get a comparison between “listreading” and “lifelike speaking”,
- to get a notion of sensitivity to background noise and OOV words,
- to have an indication of the acoustic distance between words in the proposed vocabulary.
-

In the next paragraph the test conditions in the four general tests are extensively described.

5.2 Test conditions

The technical arrangements needed to perform the tests are described in the first paragraph. Next, the speaker and the test words are discussed. The tests material was recorded in a quiet and a noisy environment, and the words were generated by two different methods. Paragraph 5.2.5 summarises the conditions applied to the four tests.

5.2.1 Technical arrangements

As the test experiments had to be reproducible, all the speech material, needed for the tests, was recorded on a digital compact cassette (DCC). The actual speech recognition tests were done afterwards in a laboratory.

Recording

The recording device was a Philips DCC-175. In all tests, an omnidirectional electret tie clip microphone: Alecto EM-110[3] with windshield was used. It was clipped to the collar of a woolen sweater at a distance of approximately 20 centimeters from the speakers mouth. The recording level of the DCC recorder was manually set to a level of -12 dB as suggested in the user manual of the DCC-175 [DCC 175 manual].

The digital speech material, of the DCC tape was copied to the harddisk of a PC. The DCC-studio software tool from Philips [DCC Studio manual] was used to label and edit the speech material. The unusefull peace's of speech were cut, and the proper speech material was labeled and ordered in tracks.

Recognition tests

The arrangement of the equipment in the recognition tests in the laboratory is shown in figure 5-1.

In the laboratory tests, the various tracks of digital speech were played back with the DCC175 connected to a PC. In this way, selected tracks of speech (i.e. utterances of the vocabulary words) could be replayed by the DCC-175, and recognized by the speech recognizer.

[3] The microphone has an frequency response of: 50-18000 Hz, a sensitivity of -52 dB, and an impedance of 1kΩ

On the recognizers side, the MSM6679 evaluation board [MSM6679-EVA-96] was connected via a serial interface to an other PC. A software package, called "Oki VRP toolkit", was used to direct the MSM6679 to start training, start recognition, show distances, etc. The analog link between the DCC-175 and the MSM6679 was accomplished using the line-out of the DCC-175 and the microphone input of the MSM6679 evaluation board. They were connected using a shielded cable with a length of approximately 1 meter.

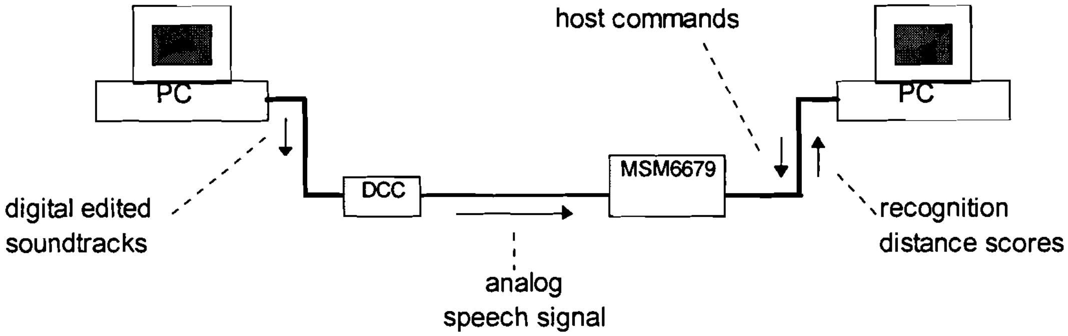


Figure 5-1, line-up of the laboratory test

In training mode, in all tests, three occurrences of each word of the vocabulary were used to train the MSM6679 (i.e. the training set). Three or more other utterances of the words (i.e. the test set) were used to test the MSM6679 in recognition mode. Upon recognition the MSM6679 was instructed (by host commands from the PC) to give distance scores, for the captured test utterance, to all templates in the vocabulary.

5.2.2 Test material

The recordings of the various utterances were split up in training material and test material; these sets were kept strictly separated. Each test has its own training-set, and its own specific test-set. In each test, the recognizer is first enrolled with the training-set, and tested (for recognition scores) with the test-set from the same test. In the OOV-test the recognizer is trained with the training-set of the KZL-test, and some "out of vocabulary" sounds are tested (for recognition scores) on this vocabulary.

The speaker

All speech material used in the test was generated by one 26 years old male speaker. The speaker is of Dutch origin, and has a southerly accent. The recordings were made in the afternoon, so the voice of the speaker was not hindered by any awakening aspects.

The vocabulary words

A standard Dutch vocabulary for environmental control was taken from a former study [Krol90]. This vocabulary was extended with a few more words. This vocabulary will probably be proposed to users in the end application. The vocabulary consists of the 27 Dutch words in the field of environmental (domestical) control; they are shown in table 5-1.

Table 5-1, words in the vocabulary

aan	bed	deur	dicht	drie
een	fout	gordijn	hoger	intercom
kanaal	lager	lamp	meer	minder
open	radio	stop	telefoon	televisie
twee	uit	ventilator	verwarming	vier
volume	wekker			

In all tests, each word is spoken at least six times; three utterances of each word are used for training, and what's left, is used for testing (recognition).

5.2.3 Test environments

The tests were conducted in two different environments; one quiet "studio-like", and the other more noisy. The four tests can be split up according to the environment as shown in table 5-2.

Table 5-2

quiet environment	noisy environment
SWL-test	KZL-test
SZL-test	OOV-test

Quiet environment

In this case the speaker is in a quiet bed/study room, sitting straight at a desk. The room is in a old house in a calm area. The only background noise may be produced by the speaker himself, or by the wind outside. The speaker wears the tie-clip microphone on his collar as described before. He sits straight-up on a chair at a wooden desk, that is faced to a wallpapered brick wall. To eliminate echoes, a damping cloth is suspended at twenty cm in front of the wall. The speaker reads his instructions from a paper lying flat on the desk; when speaking the words, he looks straight forward with his head directed to the cloth.

Noisy environment

Opposite to the latter case, this environment isn't quiet at all. The speaker is in the same house (calm area) but is now sitting relaxed on a deep couch in the living room. The living room has a parquet floor without carpets; this makes the acoustics in the room very "hollow" (much echo). The living room is situated at the street side of the building, so background noise from cars passing by, can easily occur. A radio is playing quietly at the background, and two more people (besides the speaker) are in the living room. These people are playing cards and drinking coffee (with accompanying sounds) at a distance of 5 meters. In the kitchen-part of the same room, a kettle simmers on a gas stove, and the extractor fan is running slowly. The speaker reads his instructions from a paper that he holds in his hands; again the microphone is attached to his collar.

5.2.4 word generation

Chapter 2, on user interface aspects, discussed the issue of context, semantics and emotion in the training of a SD-IWR speech recognition system. To get an indication on the influence of these aspects, two different ways to generate the command words (by the speaker) were used:

1. listreading; the words are spoken in a dull manner,
2. task driven; the words are pronounced more lifelike,

Only the SWL-test was conducted using the listreading, in the other three tests the speaker was given a task, and the words were spoken more lifelike.

Listreading

The words of the vocabulary (table 5-1) were spoken six times in random order. The six random lists were preceded by three dummy words, and at the end of the list terminated with another three dummies. In this way 6 x 27 usable utterances were acquired; three utterances of each word are used for training, and the other three are used for testing (recognition).

Task driven

In this approach, the speaker is given a message on paper, about the state of his environment (e.g. it is cold, the door is open). Next to this, he gets the task to change the state of the environment. He can change the state (i.e. turn of the fan, close the door) as demanded, by commanding the (virtual) environmental control unit with the command words of the vocabulary. In this way, aspects of context, semantics and emotion are involved in uttering the (command) words.

The tasks are presented using twenty sentences. The sentences are shown to the speaker six times, in random order. Each set of tasks is enclosed by three dummy tasks at the start, and one dummy at the end. The tasks are written down in Dutch on paper (See appendix 2) The twenty tasks generate at least one occurrence of each word of the vocabulary. By conducting the twenty tasks six times, at least six utterances of each word are obtained. Again they are strictly subdivided in a training set, and a testing set.

5.2.5 Summary

The test material was first recorded on a DCC recorder, then it was edited and labeled on a PC. The obtained utterances of the words were subdivided in a training-set and a test-set for each test. Four tests were conducted to cover task-driven and listreading speech generation, just as a quiet and a noisy environment. Table 5-3 shows the issues covered in each test.

Table 5-3, attributes of the four tests

	quiet environment	noisy environment
task-driven speech	SZL-test	KZL-test
list-reading speech	SWL-test	
OOV sounds		OOV-test

The MSM6679 was first trained, in the laboratory, with the selected training-set, next, each word in the test-set was offered to the MSM6679. The distance scores for each test word, to all words in the vocabulary were retrieved from the speech recognizer. The results of the four tests are given in the next paragraph.

5.3 Test results

This paragraph shows the results of the four tests described above. The recognition results are all based on the distance scores as being retrieved from the MSM6679. We have to bare in mind that any measured performance is the result of an interaction between the recognizer and the test material. The number of test words used in these tests, is small, and the test is performed using only one speaker. Next to this, the speech material is recorded and edited, even though the recording was done digital, and the highest care was taken in editing the speech material, it definitely affects the results. Because of all these negative aspects, we must be cautious on drawing conclusions; nevertheless, as said before the purpose of the tests is to get indications rather than scientific results.

Table 5-4, location of the various test results

test	kind of result	results in:
SWL-test	distance scores	appendix 3
SWL-test	recognition performance	appendix 4
SZL-test	distance scores	appendix 5
SZL-test	recognition performance	appendix 6
KZL-test	recognition performance	appendix 7
OOV-test	distance scores	appendix 8

The distance scores retrieved from the MSM6679 are processed in two manners; one way is to look at the distances only as a measure of distance between words, the other way is to use the distance scores as a recognition result (the word with the smallest distance to the captured utterance is the word that is recognized). Table 5-4 shows the location of the distance-scores results and the recognition performance results for all four tests.

To get an indication on **recognition performance**, the recognition performance tables should be examined. The reference templates in the vocabulary are on the vertical axis, and the test words are placed in the horizontal direction. A number n in the table on position (X,Y) should be regarded as follows: test utterance X is n times recognized as being word Y from the vocabulary.

The **distance-scores** tables show the distance between a test utterance, and all reference templates in the vocabulary. A distance n on position (X,Y) in the distance-scores tables should be considered as follows: the distance from test word X to reference template Y is n. In the appendices, three tables showing the average distance, the minimal distance and the maximal distance are given.

5.3.1 Recognition performance

The recognition performance can be expressed in the %correct measure as follows (see paragraph 5.1.1):

$$\%\textbf{correct} = \textbf{100\%} - E_{sub} - E_{del} \,.$$

In the recognition method used to obtain the recognition performance tables, there was no distance threshold for a word to be recognized; that is: even if the smallest distance to a word is, say 5000, then the utterance is still classified as being this word. In this way, there is always a word that is being recognized, and thus E_{del} has no meaning. A threshold was not used, as we didn't have a clue (at this point) on the distance scores of the MSM6679, and therefore had no idea where to put the threshold.

The %correct figures were calculated using only E_{sub} for the three tests as follows:

example: %correct for the SZL-test:

E_{sub} = the number of times a misrecognition has occurred = 27 times,

E_{sub} in percentage = 27 errors / 140 tests x 100% = 19.3%,

so %correct = 100% - 19.3% = 80.7% for the SZL-test.

If the same calculation is done for the SWL-test and the KZL-test, the results of table 5-5 are obtained.

Table 5-5, %correct scores

test	%correct	#test words
SWL-test	77,2 %	79
SZL-test	80,7 %	140
KZL-test	40,7 %	145

These %correct scores will be better, when the aspect of context is introduced. Paragraph 5.3.7 gives more details on this case.

5.3.2 Sensitivity to background noise

The %correct scores in table 5-5, from the KZL-test are far worse than those from the SZL-test. The two tests only differ in the amount of background noise in the environment. It is obvious that the background noise does influence the recognition performance. It must be said that both the training-set and the test-set from the KZL-test were recorded under severe noise circumstances.

5.3.3 Listreading versus lifelike speaking

The %correct is slightly better on lifelike speaking, than on listreading; lifelike speaking may generate more consistent speech, as the semantical, and contextual aspects are the same each time a word is spoken (in contrast to words that are read out from a list).

The test does nevertheless give no indication on the expected increased recognition performance that was discussed in chapter 3 on user interface design. The idea issued in chapter 3, was about training the recognizer with listreading words or lifelike sounding words, and recognizing only lifelike spoken words. The tests conducted here, all used the training set and the test set (i.e. the set of words that are to be recognized) from only one recording set at a time. Either the training set from listreading and the test-set from listreading were used; or the training set from lifelike speaking and test set from lifelike speaking were used. Ergo, the combination: training set from listreading and test-set from lifelike speaking is not tested.

5.3.4 Distance scores of the MSM6679

The distance scores tables, give a good indication on the range of distances that can be expected from the MSM6679. The minimal distance met, is 27 between test word "hoger" and template "hoger" in the minimal distance-scores table of the SWL-test. The maximal distance between words (not OOV sounds) is 1915 between test word "radio" and template "dicht". The maximal distance for an OOV sound to a template in the vocabulary is 3824 between OOV sound "foonzag"[4] and template "intercom", seen in the distance scores table of the OOV sounds. There seems to be no indication on the linearity of this large range, as most distance between words are between 100 and 1000. The distance of a test word to its "own" template does in most cases not exceed 300.

5.3.5 Acoustic distance between words

Concerning the "acoustic distance" between words, only the distance scores tables from the SZL-test and the SWL-test should be regarded; as these words are recorded with minimal background noise, so they give a true distance score for the sound of the words only.

The average distances in the SZL-test, and the SWL-test are smallest on the diagonal axis of the table in most cases (this could be expected). The average distance scores table from the SZL-test could best be used to give an indication on "acoustic distance" between the words, because this resembles the way the words are spoken in real usage. It seems that the words: "vier", "deur" and "meer" have a small "acoustic distance", as the distances to each other are small, and almost the same. This is also demonstrated in the recognition performance table of the SZL-test (appendix 6), where these words are mixed up four times.

There is no clear relation between the length of two words, and their distance. For example, (see average distance scores of SZL-test) test word "bed" has distance 1721 to template "aan", and distance 1244 to template "intercom", whereas test word "een" has a distance of 277 to template "telefoon". This may be, because of the linear time scaling of the speech signal, before the actual classification is performed, see paragraph 2.2 on normalization..

5.3.6 Sensitivity to OOV sounds

Most distances in the distance scores table for the OOV-test (appendix 8) are larger than 300, so they will probably not be mixed up, if a threshold of say 300 is used. The threshold can be used as a filter in classifying unknown sounds to words of the vocabulary. It can also be seen that sounds from the television (tvsoap, tvsport, tvwis) have rather high distances (>600) to the templates in the vocabulary, and even larger distances to the templates for the telephone ringing (foonhar, foonzag, foonzag) are seen. It is remarkable that three utterances of "fout" (fout1, fout2, fout3) spoken in a stressed fashion, do not have smallest distance to the template fout; the case for "stop" (stop1, stop2, stop3) is a little bit better.

[4] The OOV sounds are described in appendix 8.

5.3.7 Recognition performances using context

In chapter 3 on user interface design, the effect of syntax is broadly discussed. The tables on recognition performance can be drawn up again, taking the context into account. In this case, the recognition performance of the whole system is shown, whereas in the former case, the recognition performance of only the MSM6679 is examined. By applying the syntax as described in chapter 6, the whole vocabulary is subdivided in a lot of small vocabularies.

The recognition performance tables are now filled in another way: not the template with the smallest distance is classified as being the word recognized, but the template **in the actual vocabulary** (dependent on the state of the system) with the smallest distance is the one recognized. This yields far better recognition results; the recognition performance tables using context are shown for the SWL-test, SZL-test and KZL-test in respectively appendix 9, 10 and 11.

If the %correct numbers are calculated again, the results from table 5-6 are obtained.

Table 5-6, improved %correct using context

test	plain	using context	improvement
SWL-test	77,3 %	82.3 %	5 %
SZL-test	80,7 %	92.9 %	12.2 %
KZL-test	40,7 %	68 %	27.3 %

This table shows that the %correct measure can be improved drastically by using context (i.e. make use of syntax, as discussed in paragraph 3.1.2).

5.4 Conclusions

According to the test results discussed in the preceding paragraphs, some conclusions are drawn. The reader should bare in mind that the tests conducted are small, and just done by one speaker. Nevertheless, an indication on the recognition performance is obtained.

The **recognition performance** expressed in the %correct measure, can exceed 90% when applying context in the recognition process. With severe background noise levels, as met in the KZL-test, the %correct does not reach 70%. The SD-IWR recognition tests, are obtained with a training set of just three words, the recognition performance will increase by using more initial training templates [MSM6679-96]. The recognition performance can further be improved by template updating in operational mode of the system. When applying all these user interface aspects, the %correct in operational usage of the system should exceed 90%. Considering all this, these recognition performance results indicate that the MSM6679 could well be used in this project, to give the recognition performance required in combination with a sophisticated user interface. We must also conclude that a sophisticated user interface is absolutely necessary to compensate for the recognition errors that will occur. In the next chapter something more is said about the necessity and the implementation of the user interface shield.

Next to this most important conclusion, a few more things have come out:

- we have learned something on the distance measures of the MSM6679,
- the sensitivity to out of vocabulary sounds seems to be small,
- the MSM6679 is very sensitive to (severe) background noise, when used in both training and recognition,
- an indication on words that could be mixed up (small acoustic distance to each other) is obtained,
- and the effect of listreading versus lifelike speaking is shown.

6. System design

Before the implementation details are given, the general structure of the speech recognition unit, and the environmental control system are explained in this chapter. The system architecture of the experimental prototype is discussed, and the way the system functions is explained.

6.1 System architecture

Being in the initial, experimental phase of the project, the prototype is set up using standard development building blocks. The general structure of the system, however will be the same in the eventual integrated prototype.

Along with the demands of being small, stand-alone and low-cost, two major constraints, determine the system's structure:

- the speech interface is made for the X-10 environmental control system,
- the speech recognition sub-system is based on the MSM6679 voice recognition processor, from Oki Semiconductors.

The first constraint is the bases for this project; the project group is using the X-10 system for many years now, and there is a lot of knowledge on this particular system. The X-10 system is the defacto standard in environmental control systems nowadays, and is largely available for the public in the Netherlands. As a consequence, many disabled persons may have this system installed in their houses.

The MSM6679 satisfies the criteria required for the speech recognition system, and is selected to be the heart of the speech recognition sub-system. Chapter four extensively clarifies the reason for selecting the MSM6679.

These two choices bring about certain design issues; these have to be met in the total system design. The next paragraphs consider the consequences for the system architecture.

6.1.1 Design demands imposed by the X-10 system

X-10 is a communications protocol for remote control of electrical devices. It is designed for communications between X-10 transmitters and receivers; they communicate on a standard powernet household wiring. Receivers are generally plugged into standard electrical outlets, although some are hardwired in the device to be controlled. Transmitters send commands such as "turn on" or "dim", preceded by the identification of the receiver unit to be controlled. This message spreads over the complete electrical wiring in a building. Each receiver is set to a certain device id, and reacts only to commands addressed to it. In this way, various household appliances can be controlled individually, from one central unit (i.e. a transmitter).

As the X-10 transmitter is directly connected to the mains, there is a substantial danger of leakage currents. To give potential users the freedom of mobility, the X-10 control center should either be operated with a wireless remote control, or the system should be commanded using a speech recognition system and a microphone that is not attached to the user. In paragraph 3.4, the microphone placement issue is extensively discussed; here, the usage of a microphone not attached to the speaker, is rejected. This leaves only the option to operate the X-10 control center with a (speech-driven) remote control.

As most motor disabled mobile users, move through their house in a wheelchair, the speech recognition unit can easily be carried along, somewhere attached to the wheel chair (read paragraph 3.4). The communication between the speech-driven remote control and the X-10 control center is done using infra red code. Infra red communication is robust, and it doesn't interfere with radio-frequency (r.f.) signals. Another advantage of infra red communication over r.f., is that most television or radio sets nowadays are controlled using infra red remote controls. When using infra red communication, an extension of the system to control radio's and televisions can easily be made. This makes the system a speech-driven remote control!

6.1.2 Requirements for the MSM6679

In selecting the MSM6679 as the basis for the speech recognition system, we committed ourselves to the external peripherals needed by this chip (see paragraph 4.3.3, consequences for the project). This includes, RAM, (flash) EEPROM, an input stage to process the microphone signal, an output stage to amplify the (feedback) audio output, and a host μC (micro-controller); see [MSM6679-96] or [MSM6679-EVA-96] for detailed information.

The host μC is the most interesting part of the required periphery, as it can be used for other functions in the complete system. A simple 8-bit μC satisfies the needs to control the MSM6679 in host mode sufficiently. Next to controlling the speech recognition processor, the μC can perform the high level operations to support user interface and vocabulary management, the way this is done, is described in chapter 7: Implementation details.

6.1.3 Architecture of the speech-driven remote control

The infra red remote control requirements for the X-10 system, and the periphery needed by the MSM6679 lead to the μC being the intelligent heart of the system. The architecture of the system is shown in figure 6-1. The μC's tasks are:

- command the MSM6679 to perform the various recognition tasks,
- manage the structure of the vocabulary, using syntax,
- provide the user interface (i.e. give feedback, adaptive template updating, etc.),
- send the appropriate infra red codes to the X-10 control unit

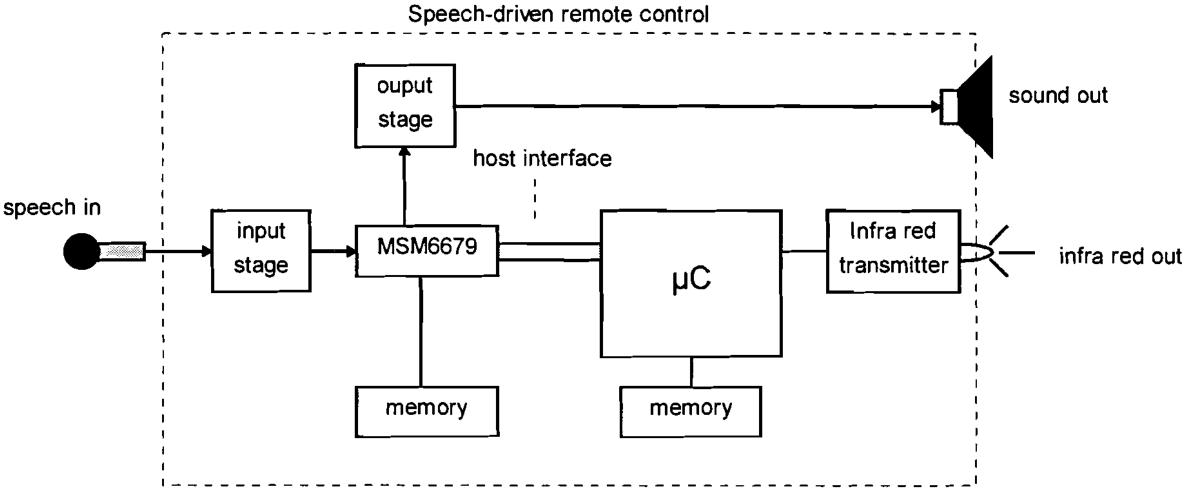


Figure 6-1, architecture of the speech-driven remote control

Implementation details of the various blocks, can be found in the chapter seven

6.1.4 Application overview

By the use of infra red codes to control the X-10 system, the speech-driven remote control for the X-10 system becomes a universal remote control. The control of appliances connected to the X-10 system is now extended to television sets, video recorders, CD-players, etc., that are controlled by a conventional (i.e. with push buttons) infra red remote control.

The system can now be incorporated in environmental control as shown in figure 6-2.

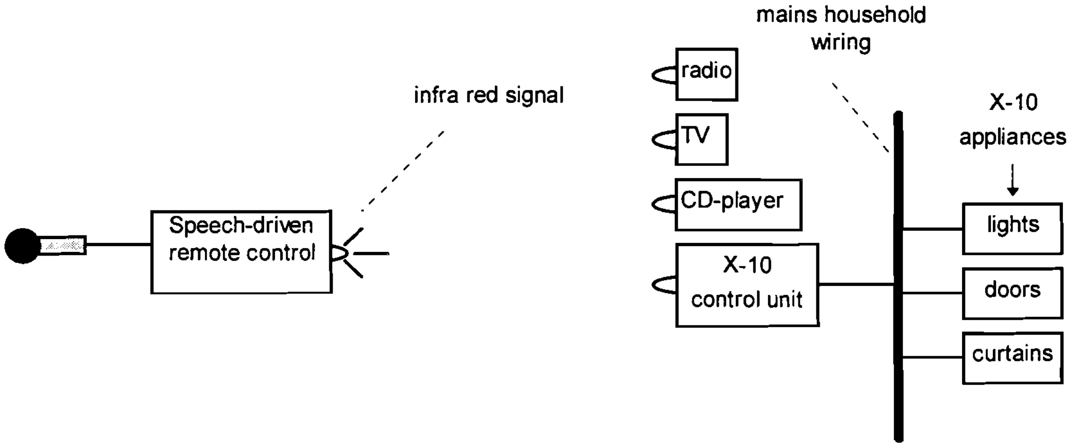


Figure 6-2, application overview

Obviously, not all infra red controlled devices use the same coding scheme for their control, yet, the various codes can be composed in software on the μC. The problem with a mobile user having more than one room in his house, can be overcome using commercially available infra red repeaters. The next paragraph discusses the users view on the device.

6.2 Functional description

An important requirement for optimal use of a speech driven environmental control system is a thorough understanding of the command vocabulary structure. That is, the user must know which commands are active, and which functions are associated with the commands at all times.

In this project, a standard set of domestical appliances has to be controlled; obviously, in the end-user application, the appliances that are controlled are different for each user. Therefore, the voice controlled interface for the (installed) environmental control system should be tailored to each user. Nevertheless, a large set of appliances and corresponding functions is used in the experimental system; see table 6-1.

The sub-appliance column is added as a way two make a distinction between more appliances of the same kind (e.g. door one, door two). The sub-appliances "channel" and "volume" should of course be regarded as a functional sub unit of the television and the radio.

Table 6-1, selected appliances and their functions

appliance	sub-appliance	function
bed	-	higher/lower
door	one, two, three, four	open/close
curtain	one, two, three, four	open/close
intercom	-	on/off
fan	-	on/off
alarm clock	-	on/off
telephone	-	on/off
light	one, two, three, four	on/off, brighten/dim
radio	channel/volume	up/down
television	channel/volume	up/down
radio	-	on/off
television	-	on/off
heater	-	turn up/down

As the system is intended for disabled persons in the Netherlands, the actual vocabulary, holding the names of the appliances, sub-appliances and functions is in Dutch language. The words used for controlling the system, are the same as used in the test in chapter five, table 5-1.

The vocabulary structure is designed in compliance with the suggestions on syntax as discussed in paragraph 3.1.2 on the number of words in the active vocabulary. This means that a command syntax is used, to reduce the number of words in each active vocabulary. Next to this, a sleep mode and operating (or command) mode is employed. As a consequence, the basic vocabulary structure looks as shown in figure 6-3. In this figure, no signs of feedback are shown; the extensive means of feedback the system uses is not shown here for clarity reasons.

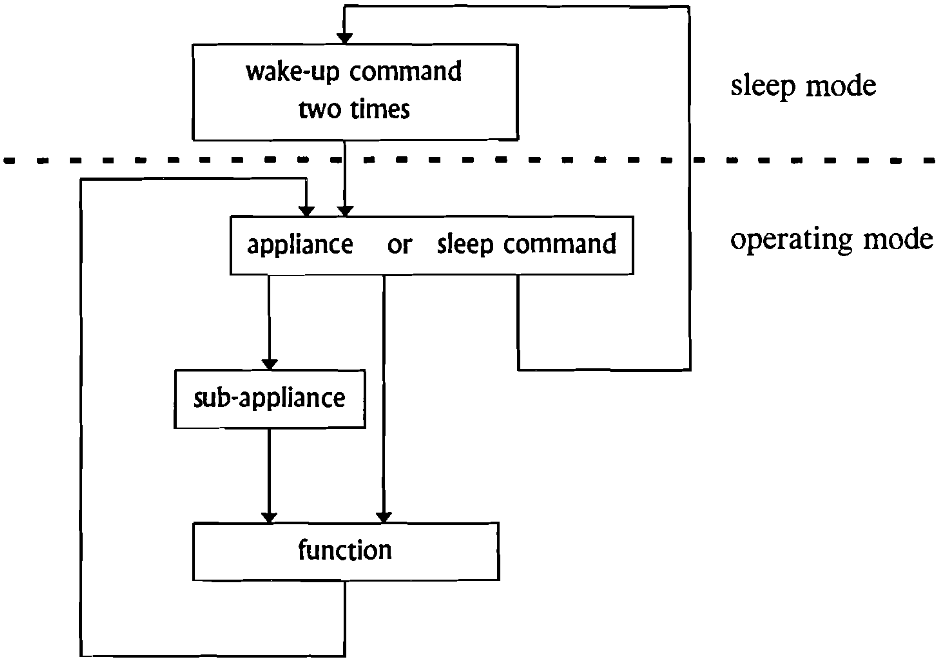


Figure 6-3, vocabulary structure

When the system is powered up, it is in sleep mode. To activate the speech recognizer, for commands used to control the appliances, the user first has to wake-up the recognition system. To do this, the word "luister" (i.e. the wake-up command) is spoken two times; after the first time the word "luister" is spoken, the speech recognizer beeps one time. This feedback is necessary to inform the user that the word was heard correctly, and that it is ready to receive the second utterance of "luister". On a mistakenly recognized occurrence of "luister", the beep from the speech recognizer functions to signal the user, that it recognized the wake-up command one time; in that case the speaker is warned not to not say "luister" again, if he was not intended to wake up the speech recognizer.

Being in operational mode, the user can give commands to control the appliances connected to the environmental system. A command is always started by speaking the name of the appliance to control. The name of the appliance is echoed by the speech recognition system, so the user gets feedback on the recognized word. Next to this, the echo is used to time the input of the user (see paragraph 2.2 on feedback). After the echo, the user can either speak a sub-appliance (see table 6-1), or a function command, dependent on the appliance he/she wants to control. The name of the appliance and (the sub-appliance, and) the function is echoed one more time, for the same reason as above. At the end, the user can confirm this complete echoed command by saying "ja" (i.e. yes) or reject the command by saying "nee" (i.e. no). The dialog between the user and the speech recognizer for each command, in command mode is shown in appendix 12. The square boxes in this figure hold the words that are together in one active sub-vocabulary. In appendix 12 the largest sub-vocabulary, constituted of all appliance names, the sleep command: "slapen", and the cancel command: "fout", is not shown.

After a whole command is completely recognized and confirmed, the system sends the appropriate infra red code to the X-10 IR-receiver, and the appliance is activated accordingly; meanwhile the MSM6679 says: "working" (speech synthesis), to indicate that the system is performing the required action. The speech recognition system returns to the command state of the vocabulary structure (figure 6-3), and the next command can be issued.

If the user doesn't want to control any more appliances, he/she can put the system in sleep mode. This is done by speaking the sleep command "slapen" one time. The speech recognizer reacts with a double beep, to indicate, it's going to sleep mode.

In an erroneous situation in the dialog (i.e. the recognizer has incorrectly recognized a command), the user can always go back to begin of the command mode by saying "fout" (i.e. cancel). In each state of the dialog, this command may be used to get out of an unwanted situation. When the command "fout" is recognized, a special "beep" is fed back to the user, to indicate that something went wrong, and no infra red code was send, so no appliance was activated. At this point, the user can proceed by giving a new command, or go to sleep mode as usual.

A special feature is added, for repeatedly applying the same function to an appliance. For example, if the user wants to brighten a certain light (say, light one: "lamp" "een"), the dialog is as given in table 6-2. Being in command mode, the user starts the command by speaking the appliance "lamp" (row 1 in table 6-2). Next, the sub-appliance "een" is spoken, the speech recognizer echoes: "lamp" "een" (row two). Then, the function is commanded: "meer", and the complete command is echoed (row three). The user confirms the given command by saying "ja", and the command is executed by the environmental control system; that is, the light is brightened one discrete visual step. Up to here, the dialog is standard, but once again the speech recognition system echoes the function command: "meer" (row five). The user can either confirm the function to be applied once again (i.e. brighten the light one step more) by saying "ja" (row six), or not (the light is yet bright enough) by saying "nee" (row eight). In all cases, after the command is executed (i.e. the infra red code is sent), the speech recognizer says "completed", and the system is back in the initial command mode.

Table 6-2, repeating a function command

	user says:	feedback from the speech recognizer:
1 start of dialog	“lamp”	“lamp”
2	“een”	“lamp” “een”
3	“meer”	“lamp” “een” “meer”
4 confirmed	“ja”	(command is executed)
5		“meer”
6 confirmed	“ja”	(command is executed once again)
7		“meer”
8 affirmed	“nee”	“completed”
9 end of dialog		

The special feature described in the latter is applied for the functions: up/down, brighten/dim, turn up/turn down, and higher/lower; in Dutch: “hoger”/”lager” and “meer”/”minder”. The repeat loop for these functions is clearly shown in appendix 12. in the command trees for: “tv”-”radio”, “bed”-”verwarming” and “lamp”.

7. Implementation details

This chapter describes the way the experimental prototype of the speech driven remote control is implemented in detail. Note, that this first experimental implementation is set up using standard components to speed up the prototyping. Most components used are overdimensioned, and in this stage, aspects of physical dimensions and power consumption are not taken into account yet. The main purpose of this first prototype is to look at the possibilities to build a speech driven environmental control system using standard, low cost (of the shelf) components.

7.1 The µC

The μC used is a derivative of the INTEL 8051 8-bit microcontroller family. The 80C32 member was chosen, for the three timers it inhibits (to compose the infra red code, and for timing of the serial port). This ordinary μC consists of 256 bytes of internal RAM, a serial port, three internal timers/counters and four 8-bit bi-directional ports. Extended information on the 80C32 μC can be found in the Philips databook [IC20-95].

To keep up with the MSM6679 running on 32 MHz, the 12 MHz version of this μC was chosen. An external flash EEPROM of 64 kB was used to hold the program code for the μC; the advantage in using a flash EEPROM is the short time to load and erase the contents of this memory (rapid prototyping). Two 8-bit ports were used as a multiplexed data/address bus for data transfer between the μC and the memory. The third bi-directional port was used to control the infra red transmitter.

7.2 Communication with the MSM6679

To get aquainted with the MSM6679 voice recognition processor, an evaluation board [MSM6679-EVA-96] was used. The evaluation board is equipped with audio input and output stages, power, reset and clock circuitry, 128 kB of RAM and Flash EEPROM, and an Oki MSM66P54 speech synthesis chip. Because the evaluation board was used for testing the MSM6679 (see paragraph 5.2.1) and in the prototype, it should be able to communicate with both a PC (testing), and the μC (prototype). Therefore, the communication was done over a detachable serial link; the standard RS-232 serial interface protocol.

To accomplish this, the μC was equipped with a MAX232 serial interface communication chip; this chip was connected to the serial port of the μC. In this way the slave mode API (see paragraph 4.3.2) was used to command the speech recognition processor over the serial interface.

According to the slave-mode API [MSM6679-96] the host commands consisted of four bytes, representing the ASCII value of the opcodes. The various opcodes for the MSM6679 are extensively described in [MSM6679-96]. In general they consist of four characters forming a 4 digit hexadecimal number like F340. The opcodes were sent over the serial line one byte at a time. Each received byte was immediately echoed by the MSM6679, so this could be used as a way of communication error detection.

7.3 Infra red communication

As described in chapter 6, the communication between the speech recognizer and the X-10 environmental control system is done using infra red (IR) signals. The speech driven remote control holds an IR transmitter, to communicate with an existing IR receiver for the X-10 system.

7.3.1 The IR receiver

The infra red receiver is a commercially available device, that can receive infra red codes, and transmit the appropriate X-10 codes over the mains. The trade mark name of the infra red receiver is "Command Center URC 3000", it is manufactured by Universal Electronics Inc., but retailed by "One For All" under the name IR543 (further, it will be referred to as IR543).

The IR543 is designed to be operated with a "One For All" remote control; it can also directly operate eight X-10 devices with the buttons on top of the device. In conjunction with a infra red remote control, the IR543 can control all X-10 devices (i.e. device one to sixteen, for sixteen different house codes) with all functions (i.e. on/off, brighten/dim).

7.3.2 IR Transmission Protocol

The format of the infra red code used by the IR543 is not public available. The format of these infra red codes, were examined with a IR sensitive photo diode attached to an oscilloscope. The IR signals to control the IR543 were sent by the accompanying "One For All" remote control (having conventional push buttons).

The infra red transmission format is shown in Figure 7-1. It consists of a start code, 5 bits of complemented data, the original data (K4,...,K0), and a stop code. The start code is represented by a 4 ms burst of approximately 40 kHz, followed by 4 ms of an absence of 40 kHz (i.e. a logical one). Each data cell is 8 ms wide. A logical one is represented by a 4 ms burst of 40 kHz, while a logical 0 is represented by a 1.2 ms burst of 40 kHz. The two times five data bits are then followed by a stop code of a 12 ms long burst of the carrier of 40 kHz.

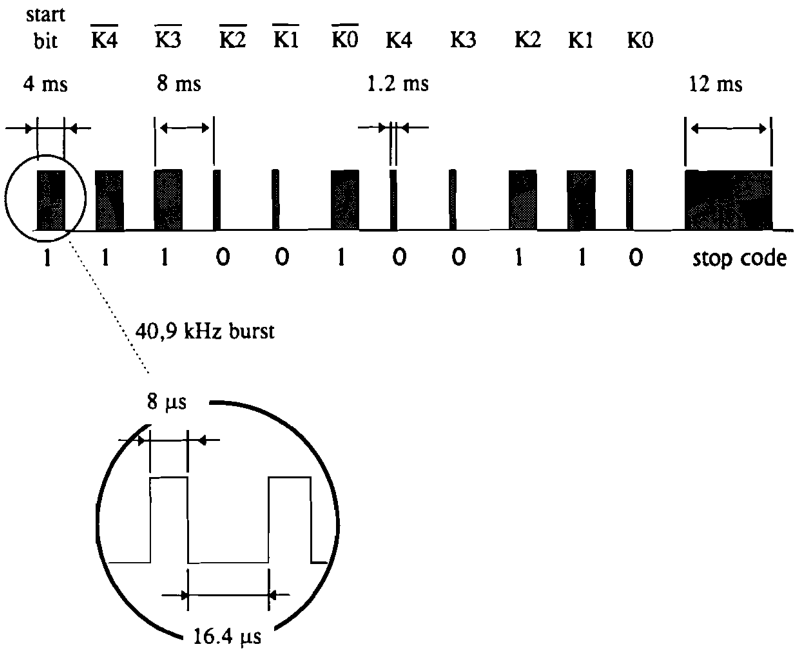


Figure 7-1, infra red transmission format

The five data bits, hold the information for X-10 control, i.e. a device number (one to sixteen) or an X-10 function. An X-10 device can be controlled, by first sending the infra red code word to select one out of sixteen devices, next another code word is sent representing the function to be applied. Appendix 13 shows the data bits used for each X-10 device and function. In the case were the function for a specific device is repeated (see paragraph 6.2), only the function code word is repeatedly sent, as the X-10 system holds the device id of the last addressed appliance.

7.3.3 The infra red transmitter

The infra red codes described above, are composed in software, on the μC; the actual burst of 40 kHz is generated by a burst generator (i.e. the infra red transmitter). An output line of port 1 of the μC is used to drive the infra red transmitter; i.e. turn the transmitter on and off. (see appendix 14 for the electric circuit)

The infra red transmitter is composed of an external timer IC: NE555, an IR-LED driver unit, and an interface circuit between the CMOS based μC and the TTL-level NE555. The NE555 general purpose timer/counter IC, is used as a free running astable multivibrator generating a two level output, that is the burst signal of 40 kHz. The shape of the burst is shown in figure 7-1; in one period, the signal is high for 8 μs, and low for 16.4 μs. The trim resistors around the NE555 are used to fine-tune the burst exactly to 40.9 kHz, with a dutycycle as described above.

The astable multivibrator is switched on and off by pulling the reset pin of the NE555 low or high. This is done by an extra transistor stage that couples the CMOS output pin of port one of the μC, to the TTL reset input of the timer IC.

The output line of the NE555 switches the transistor of the infra red LED driver stage. The transistor drives three infra red LED's, to get a widespread reach of the infra red signal.

7.4 Software management

7.4.1 Nametag and vocabulary storage

The MSM6679 can play back 61 nametags (i.e. short pieces of recorded speech), for feedback usage. The memory size of the EEPROM determines the duration of each nametag. The evaluation board holds a flash EEPROM of 128 kB, that can hold 27 seconds of speech. Before a nametag can be recorded, its duration must be set; note that all nametags must have the same length for the MSM6679. In the actual configuration, 25 nametags are recorded with a duration of one second. In the end user system, the nametags should be recorded using a professional speaker in a studio. Next to feedback by nametags, the speech synthesizer of the evaluation board, and the build-in tone generator are used to give feedback (e.g. “working”, “beep”).

The vocabulary of the words to be recognized (see table 5-1) are stored in the flash EEPROM of the MSM6679 evaluation board. At this moment, the words are only used for testing the software and hardware of the system, and not for actual recognition purposes. That's why the words were trained in a laboratory, and no effort was done yet, to get a high recognition performance as described in chapter three, on user interface design. The speaker dependent vocabulary is transferred to RAM when the system is powered up, and can be saved to flash EEPROM and to a file on a PC (i.e. upload data) any time.

In fact the MSM6679 uses only one large vocabulary, but the structure of the program in the μC provides the sophisticated use of syntax. The way this is done, is discussed in the next paragraph.

7.4.2 Program structure

The software program is written in 8051 assembly language and occupies 2 kB of memory in the μC's program memory. The program is constructed in a bottom up manner. That is; first the low-level subroutines were written, to provide for the basic steps in:

- communicating with the MSM6679 over the serial port (e.g. send a byte),
- timing the infra red transmitter in sending logical one's and zeroes (e.g. transmit a zero).

The higher level routines, (e.g. send MSM6679's commands, send the infra red code word to dim a X-10 device) make use of the low level subroutines; in this way the complete software program was constructed. In the highest level, the software program acts as a state machine. The structure of the state machine is roughly the same as the one discussed in paragraph 6.2 (figure 6-3). Appendix 15 shows the state machine in more detail.

The **feedback** given by the system in all states or transitions of the state machine is given by using nametags, tones, or synthesized speech (see paragraph 4.3.1). The feedback of the system is shown in a rounded box in appendices 12 and 15.

The **recognition** of a word, is based on the selection of the word (in the active vocabulary) that has minimal distance to the captured utterance; see paragraph 5.3.7 on using context and paragraph 6.2 on the words that are in the same active vocabulary. A threshold in this selection is not applied yet.

The program starts with initializing the serial port of the μC, next, the MSM6679 is initialized to function in speaker dependent recognition mode with the appropriate vocabulary (the vocabulary is restored from flash EEPROM to RAM memory). After the initialization, the state machine is triggered by utterances of the speaker, to go from one state to another. The state machine goes to command mode, after the wake-up command "luister" is heard two times. To inform the user about the systems state a "beep" is sounded two times.

Being in command mode, the speaker can either say "slapen" (i.e. the sleep command) or the name of an appliance. If an appliance is heard, the state machine parses the command tree of the various appliances as shown in appendix 12. At the end of such a command tree, the infra red code for the parsed X-10 command is constructed, and sent by the infra red transmitter (the MSM6679 says: "working").

If the infra red code words are sent, the state-machine goes back to the command mode state, to listen for the next command to be uttered (i.e. an appliance or the sleep command). In each state of the state-machine, the user can go back to the command mode state by uttering the cancel command: "fout"!

8. Conclusions and suggestions

This project is the first step in the development of an end user speech driven interface to the X-10 environmental control system. The eventual product must be optimized for energy consumption, size and user friendliness. It should not only be an extension to the range of input-devices for the X-10 system, but a device that enables motor disabled persons, who are not able to use an environmental control system because of their handicap, to have control over their immediate domestic environment.

In this project the basis for the implementation of such a device is set. Being in the initial stage of the project, the speech driven interface to the X-10 system is not yet optimized; therefore several useful suggestions for further work are done, at the end of this chapter.

Nevertheless, the goal of this project was reached well; i.e. investigate the possibilities of present speech recognition technology, for usage in a low cost, stand-alone speech recognition system for environmental control, by motor disabled persons. Next to this, a first prototype with low cost, “of the shelf“ components should be implemented. The next paragraph discusses the results from this project.

8.1 Conclusions

Looking at the aim of this project, the most important conclusion is as follows:

> **It is possible, to build a low cost, stand-alone speech recognition system with the current speech recognition technology. A first prototype that controls the X-10 system by voice command is implemented.**

Several findings, encountered during the project, led to this main result. These sub-conclusions, forming an important part of the feasibility study, are enumerated below.

- An extensive field study led to the selection of the MSM6679 voice recognition processor from Oki Semiconductors. This stand-alone, host driven, dedicated digital signal processor is used for speaker dependent isolated word recognition of 32 words out of the environmental control domain.
- Various recognition performance tests of the MSM6679 indicate that this speech recognition chip can well be used to control the X-10 system by voice commands. In cooperation with a well designed vocabulary structure, the %correct performance measure can exceed 92 %. The test results confirm the idea that the use of context increases the recognition performance.

- A first experimental prototype, based on the MSM6679 is built. The prototype is portable, and not physically attached (i.e. electrical save) to the X-10 system. To accomplish this, the system is fit up with an infra red transmitter. The IR transmission protocol of a commercially available IR-transmitter/receiver set (for the X-10 system) is applied to the IR-transmitter. In this way, the X-10 system (equipped with the IR-receiver) is wirelessly controlled by a speech-driven remote control. The incorporation of the flexible IR-transmitter extends the functionality of the device, to control other devices such as: TV, radio, CD- player, etc., that are controlled with conventional IR remote controls.
- A sophisticated command dialog, with extensive feedback mechanisms is incorporated in this experimental prototype. Together with the aspects of vocabulary management, various user interface design issues for a voice commanded environmental control system are proposed.

8.2 Suggestions for further work

Both in ergonomics, as well as in electrical engineering, further development of the experimental prototype is necessary. The bulk of effort has to be spent in optimizing the system. Nevertheless, there are some specific design aspects, I would like to propose; they are given below.

Human factors issues

- The recognition of a word is now based on the distance scores; i.e. the reference template in the active vocabulary that has minimal distance to the captured utterance is classified as the word that is spoken. Ergo, there is always a word recognized, even if this minimal distance were very large. To prevent too many substitution errors, a threshold should be applied in the classification.
- The nametags used for feedback, are recorded poorly. They should be spoken by a professional speaker, in a studio. The actual prototype, uses synthesized English spoken words for feedback, these should be replaced by nametags spoken in Dutch language.
- Next to auditive feedback, the system could be expanded for deaf persons by giving additional visual feedback (e.g. with a small LCD, or LED's). An experienced (not deaf) user could turn off the audible responses, and just inspect the display for feedback concerns; note that audible responses can become intrusive when they are constantly heard!
- The proposed initial enrollment strategy (give the speaker virtual tasks) should be brought in practice; i.e. an explicit training strategy should be developed. The MSM6679 offers the possibility for adaptive template updating; the confirmation of a word that will be updated can be done using speaker independent recognition.

Engineering aspects

The experimental prototype is developed using standard over-dimensioned building blocks; functionality was the first design criterion. Obviously, the end user product will not include the MSM6679 evaluation board, and probably the speech synthesis chip is omitted. The complete design has to be optimized for costs and energy consumption; nevertheless, there are some specific suggestions that will be useful in the design.

- The carrier of the IR signal (i.e. the 40kHz burst) is generated by an external timer/counter IC. It should be possible to construct the burst signal on the microcontroller, using an internal timer. This would reduce the energy consumption of the system, as a complete IC is omitted in the design.
- The external program-memory (i.e. EPROM) for the μC can be replaced by a μC with a small internal EPROM memory, for the amount of program code is small (< 2 kB).
- Because all signal processing and classification, to perform speech recognition, is done on the MSM6679, the host μC isn't occupied that much. The simple 80C32 μC (just as most simple 8-bit μC's) incorporates sufficient residual capacity (i.e. output pins and processing time), to provide for an extra, visual means of feedback.

Bibliography

[Atal95] Atal, B.S.
"Speech technology in 2001: New research directions."
In: proc. of Natl. Acad. Sci.: Human-machine Communication by voice, Irvine, February 8-9 1993, Vol. 92, pp. 10046-10051, October 1995

[Bloks87] Bloks, R.H.J.
"Een spraakherkenner voor het Busch Timac X-10 afstandsbedieningssysteem."
Afstudeerrapport, Vakgroep Medische Elektrotechniek, Faculteit der Elektrotechniek, Technische Universiteit Eindhoven, Eindhoven, Nederland, 1987

[Bosch85] Bosch, H.
"Een goedkope spraakherkenner voor toepassing in hulpmiddelen voor motorisch gehandicapten."
Afstudeerrapport, Vakgroep Medische Elektrotechniek, Afdeling Elektrotechniek, Technische Hogeschool Eindhoven, Nederland, 1985

[Bradford95] Bradford, J.H.
"The human factors of speech based interfaces, a research agenda."
SIGCHI bulletin, Vol. 27(1995), No. 2, pp. 61-67

[Cox90] Cox, S.J.
"Hidden Markov models for automatic speech recognition: theorie and application",
In: Speech and language processing ed. by Wheddon, C. and R. Lingard, London, Chapman and Hall, 1990, pp 209-230

[Custers97] Custers, R.J.G.
"Manual Spraakwater; Syntaxis and semantics of the mnemonic language Spraakwater, for the Oki MSM6679 Voice Recognition Processor"
Intern rapport [97EME-02], sectie Medische Elektrotechniek, vakgroep Meet- en Besturings Systemen, faculteit der Elektrotechniek, Technische Universiteit Eindhoven, Nederland, 1997

[DCC 175 manual] "Philips DCC 175 Digital Compact Cassette Recorder Gebruiksaanwijzing", Philips electronics N.V., Eindhoven, The Netherlands, 1995

[DCC-Studio manual] "Philips DCC-Studio gebruiksaanwijzing"
Philips Electronics N.V. Eindhoven, The Netherlands, 1995

[Green83] Green, T.R.G. and S.J. Payne, D.L. Morrison, A. Shaw
"Friendly interfacing to simple speech recognisers."
Behaviour and Inf. Tech., Vol. 2(1983), pp 23-38

[IC20-95] "Philips data handbook IC20, 80C51 based 8-bit microcontrollers"
Philips Semiconductors, March 1995

[Johnston96] Johnston, R.D.
"Are speech recognisers still 98% accurate, or has the time come to repeal 'Hide's Law' ?"
BT Technology Journal, Vol: 14, No. 1 Januari 1996, pp. 165-176.
BT lab. UK, Januari 1996.

[Krol90] Krol van der, R.C.P.
"Demonstratiemodel van een spraakherkenner ten behoeve van omgevingsbesturing door gehandicapten."
Afstudeerrapport, vakgroep Medische Elektrotechniek, Faculteit der Elektrotechniek, Technische Universiteit Eindhoven, Nederland, 1990

[McCauley84] McCauley, M.E.
"Human factors in voice technology."
Human factors review 1984, Ed. By Muckler, F.A.; managing ed. Neal, A.S.; prod. Ed. Strother, L., Santa Monica 1984, pp. 131-161

[MSM6679-96] MSM6679 Voice Recognition Processor, datasheet
Sunnyvale, USA: Oki Semiconductor, Februari 1996

[MSM6679-EVA-96] EVA KT 6679-2 Voice Recognition Processor Evaluation Board
Sunnyvale, USA: Oki Semiconductor, March 1996

[Noyes89] Noyes, J.M., and R. Haigh, and A.F. Starr
"Automatic speech recognition for disabled people."
Applied Ergonomics, Vol. 20(1989), No. 4, December 1989, pp. 293-298

[Neutelings87] Neutelings, P.
"Spraakherkenning met de SP1000 spraakanalyse chip."
Afstudeerrapport, Vakgroep Medische Elektrotechniek, Faculteit der Elektrotechniek, Technische Universiteit Eindhoven, Nederland, 1987

[Poll91] Poll, L.H.D.
"Een spraakgestuurde afstandbediening."
Afstudeerrapport, Vakgroep Medische Elektrotechniek, Faculteit der Elektrotechniek, Technische Universiteit Eindhoven, Nederland, 1991

[Rabiner93] Rabiner, R. and B. Juang
"Fundamentals of speech recognition".
PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993
Prentice Hall Signal Processing series

[Rowden92] Rowden, C.
"Analysis."
In: Speech Processing ed. By Chris Rowden, London, McGraw-Hill, 1992, pp 35-73, the Essex series in telecommunications and information systems.

Appendix 1

Table of speech recognition IC's

	DVC306	D6106	HM2007	MSM6679	RSC164
Manufacturer	DSP Communications inc.	DSP Communications inc.	Hualon Microelectronics	Oki Semiconductors	Sensory Circuits inc.
Introduction year	1995	1995	1995	USA 1996 EUR 1997	1995
μC needed ?	Yes	Yes	No	Yes	No
maximum external memory	1 MB SRAM 1 MB ROM	128 kB SRAM 128 kB ROM	8 kB SRAM	256 kB SRAM 256 kB ROM	64 kB SRAM 64 kB ROM
Price, large quantities	$10	$8	$16	$25	$5
SD recognition	Yes	Yes	Yes	Yes	Yes
SI recognition	Yes	No	No	Yes	Yes
#words SD	8 sets of 128 words	16-128	40 of 0.9 secs. 20 of 1.9 secs.	3 sets of 61 words	2-20 each set
#words SI	128	-	-	standard 20	2-10 each set
speech synthesis	Yes	Yes	No	Not on-chip	Yes
Record/playback	Yes	Yes	No	Yes	Yes
CODEC needed	Yes	Yes	No	No	No
response time SD/SI (in ms)	300/500	700/-	300/-	200/200	?
Clock speed (MHz)	32	29	?	32	14,3
Power consumption min/max (mW)	50/260	300	30/75	0.05/50	0.05/25
evaluation set	Yes	Yes	Yes	Yes	Yes
special	DTMF dial tones		direct microphone input	DTMF dial tones host interface protocol interface to Oki speech synthesis chips	four channel music synthesis direct speaker output
remarks	successor of D6106			available in Europe	based on neural network

Appendix 2

Task sentences for generating words

	task sentences (in random order)
dummy	God wat is het hier koud, doe deur 3 eens dicht !
dummy	Het tocht hier behoorlijk, doe de vantilator maar uit hoor.
dummy	Nu wil ik naar Bonanza kijken, zet de TV eens op kanaal 7, hij staat nu op kanaal
	Het wordt al donker, en ierdereen kijkt naar binnen, doe het gordijn maar dicht !
	Het wordt al donker, en ierdereen kijkt naar binnen, doe gordijn twee maar dicht
	Zo we gaan maar eens, maar hoe gaat de deur open?
	Het bed gaat nu langzaam omhoog, zeg maar stop als het op de goede hoogte staat.
	Hoe kun je nu in zo n hoog bed stappen, je kunt het toch wel wat lager zetten ?"
	Doe die deur eens open, ik geloof dat het deur 1 is.
	Zet de tv eens op kanaal drie !
	Zet het volume van de radio eens wat hoger.
	Nee, knuppel je moet niet de deur, maar het raam openen, zeg fout !"
	Ik kan het niet goed lezen, doe de lamp eens aan.
	Bonanza is er op, zet de TV eens aan:
	Wat blaast hier zo, zet de ventilator eens uit."
	Wat brandt lamp 4 toch fel, zet hem eens wat minder !
	Wat is het weer koud hier, zet de verwarming eens wat hoger.
	Goh wat is het hier stil, mag de radio aan?
	Hoor je niet dat de telefoon gaat, pak hem eens op.
	Zet die bak toch eens uit, heb je niks beters te doen ?
	Er is iemand aan de deur, luister eens via de intercom !
	Je moet morgen vroeg opstaan, zet je wekker even ja?
	Ook al heb je de gordijn geopend, het is hier toch nog erg donker, zet lamp 4 eens wat meer.
dummy	God wat is het hier koud, doe deur 3 eens dicht !

Appendix 3

Distance scores tables for the SWL-test

Average distances SWL-test

reference words in the vocabulary   spoken test utterances ---→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	112	813	908	1233	797	361	587	1170	775	1140	822	926	1084	961	1436	568	792	839	1107	1386	1018	385	1153	547	1447	815	1016
bed	1114	268	401	393	504	713	1093	723	517	760	812	643	906	424	601	827	1113	810	624	838	770	982	822	663	692	570	506
deur	1133	512	233	493	679	662	1213	1175	653	1276	1201	467	891	293	621	1045	1097	1307	908	862	804	1113	1334	768	599	737	535
dicht	1286	348	456	268	499	770	1137	765	641	1029	1131	718	891	506	694	935	1115	991	709	698	685	1000	1082	817	771	641	625
drie	821	416	601	598	213	468	564	701	621	1172	1000	732	505	622	1011	787	546	794	581	358	153	615	1259	425	723	328	833
één	377	497	582	821	378	133	490	675	466	794	682	504	585	498	768	380	454	533	562	623	344	413	824	337	755	345	572
fout	236	568	828	967	485	255	232	865	566	1006	703	884	720	835	1277	540	570	364	825	971	528	313	1025	405	1087	497	829
gordijn	1030	685	894	983	638	599	881	261	470	532	834	524	820	644	593	363	688	702	375	518	660	915	563	600	739	423	483
hoger	672	440	425	690	415	356	637	383	81	560	411	377	690	310	362	338	416	533	328	638	677	678	596	369	512	256	280
interc.	1030	1092	1001	1251	937	709	1231	706	584	54	698	734	1081	752	747	375	971	710	532	960	1100	1250	70	594	978	576	516
kanaal	626	571	734	1038	552	539	689	616	363	605	120	786	957	732	826	578	526	580	458	981	860	742	681	384	937	417	718
lager	846	521	350	807	583	486	909	566	296	672	749	86	567	215	308	489	590	988	421	557	614	965	736	426	387	337	256
lamp	459	437	487	780	431	336	501	574	430	505	666	392	241	379	704	391	429	606	401	533	338	582	587	233	597	273	494
meer	1166	523	309	706	627	689	1092	670	294	818	826	247	677	104	266	695	718	1071	476	521	688	1135	902	513	244	336	292
minder	1274	675	484	902	762	751	1211	561	271	764	790	259	828	250	63	607	757	1048	488	629	869	1187	814	703	368	486	245
open	898	1062	998	1193	918	615	1145	575	552	374	960	599	1066	625	765	164	750	963	538	765	1016	957	367	640	984	525	508
radio	620	798	852	1147	635	476	595	513	459	950	631	701	697	672	872	411	131	853	466	602	604	678	999	372	870	306	804
stop	533	646	886	1047	558	434	469	872	524	728	568	914	798	848	1057	688	848	207	772	1137	739	640	771	518	1017	588	703
telef	911	628	622	863	513	472	894	358	323	318	522	474	751	440	484	315	456	697	188	413	555	949	369	360	573	194	445
telev.	956	584	547	646	375	444	867	394	422	575	812	468	636	378	514	375	472	832	282	146	341	808	636	371	498	210	488
twee	510	478	642	716	308	248	372	674	572	929	818	696	477	565	1018	498	394	574	566	390	147	367	979	287	748	277	749
uit	371	673	722	887	508	230	604	627	421	374	572	587	745	564	834	170	562	473	541	763	636	450	379	322	915	381	484
ventil.	795	873	803	1054	719	510	966	586	443	97	565	604	912	596	690	280	742	584	427	759	857	996	118	370	817	379	450
verwar.	485	555	564	857	398	284	536	452	298	408	470	489	638	469	742	287	393	518	337	504	503	639	461	117	631	164	509
vier	1138	646	365	751	622	565	1009	769	359	781	946	317	689	196	292	716	842	843	559	531	645	1177	844	554	242	433	254
volume	726	465	620	788	399	396	579	415	314	651	574	572	653	430	701	406	361	529	312	411	366	679	703	257	568	94	562
wekker	1098	697	583	888	682	582	1017	537	219	555	795	244	817	295	234	444	827	738	492	755	871	1055	565	679	455	487	68
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	3	4	3	3	4	3	3	3	3	3	2	3	2	4	2	3	2	2	3	2	3	4	3	2	4	3	3

Minimal distances SWL-test

reference words in the vocabulary ↓ spoken test utterances →

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	81	625	764	1162	604	245	446	1074	732	1058	633	883	890	794	1317	475	562	672	985	1304	875	277	1096	452	1247	769	864
bed	977	155	344	265	413	615	920	654	432	676	710	511	598	371	496	810	966	696	471	818	628	861	768	525	540	446	404
deur	1009	343	115	332	563	537	1142	948	640	1174	1109	409	577	249	516	966	937	1144	725	808	646	1072	1292	617	417	607	462
dicht	1108	237	334	180	416	615	917	721	521	931	1042	613	650	455	575	900	997	867	571	658	597	868	1001	625	619	534	449
drie	790	343	552	410	86	294	512	570	562	1049	966	712	419	592	950	731	481	731	335	346	130	565	1210	359	646	287	716
één	319	404	506	734	229	95	389	603	400	699	566	476	448	415	660	374	270	419	423	587	220	301	812	323	643	273	479
fout	218	454	749	816	342	204	114	821	499	915	555	816	665	722	1157	506	436	308	690	935	483	239	1002	344	941	419	762
gordijn	957	653	826	879	588	518	737	147	397	491	767	456	779	606	583	297	560	618	271	502	604	779	471	490	568	362	311
hoger	569	384	415	670	311	297	503	347	27	485	341	335	589	261	324	276	352	511	169	610	507	605	476	297	405	253	224
interc.	947	1038	880	973	726	640	1098	643	552	53	567	637	837	722	716	356	823	544	301	951	964	1077	66	549	713	527	479
kanaal	561	455	615	917	462	499	565	530	295	518	90	698	831	670	803	552	443	538	317	972	701	649	621	374	831	374	654
lager	791	450	292	693	475	398	800	467	278	594	704	32	375	176	299	428	452	830	313	536	481	887	728	358	341	285	163
lamp	413	353	405	647	371	281	385	505	420	415	645	310	86	350	658	316	344	453	308	471	225	460	492	232	470	219	449
meer	1087	435	200	646	562	562	1014	467	250	724	820	216	448	82	266	646	625	1001	371	504	583	1028	858	347	157	294	267
minder	1138	602	342	790	636	645	1152	469	221	684	742	204	660	232	50	540	636	955	395	599	717	1112	751	522	267	441	183
open	868	975	915	1024	763	562	1033	437	510	312	834	555	868	574	749	116	561	855	435	665	944	849	286	553	706	504	421
radio	587	666	775	924	596	404	514	364	390	845	598	650	696	595	864	405	101	678	407	577	447	613	929	336	840	248	745
stop	427	567	822	1018	367	381	303	840	469	643	432	852	670	810	914	659	749	79	544	1120	630	537	737	470	742	498	633
telef	878	608	565	829	421	348	855	260	242	257	470	413	633	417	467	287	344	684	117	401	463	890	311	258	503	186	337
telev.	897	535	486	617	280	273	780	350	366	494	802	440	520	347	514	352	354	755	132	100	304	747	612	208	479	175	372
twee	505	363	582	484	212	156	281	624	525	821	798	657	420	469	958	475	277	516	310	355	135	360	936	235	717	220	666
uit	337	584	674	827	326	176	526	543	384	334	441	540	548	471	746	148	373	332	432	702	529	394	366	322	710	343	414
ventil.	739	817	747	862	536	474	842	549	407	84	442	521	689	565	666	260	612	452	244	739	736	851	96	329	587	345	394
verwar.	480	506	539	805	305	238	449	362	260	350	417	451	525	425	728	256	316	457	269	503	402	556	428	106	529	160	449
vier	1016	534	206	656	474	483	912	492	347	698	869	294	501	120	246	677	745	707	427	508	535	1093	809	395	90	366	243
volume	700	393	543	580	328	300	492	335	225	534	514	552	558	388	688	379	293	469	201	379	326	644	638	171	451	63	467
wekker	983	617	386	711	572	540	870	508	198	496	697	199	631	234	169	406	690	640	338	720	737	959	494	574	280	423	34
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	3	4	3	3	4	3	3	3	3	3	2	3	2	4	2	3	2	2	3	2	3	4	3	2	4	3	3

Maximal distances SWL-test

reference words in the vocabulary ↓ spoken test utterances ---→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	151	1052	1093	1337	989	426	783	1230	797	1263	1011	1000	1279	1049	1556	666	1022	1006	1273	1469	1101	465	1188	643	1714	859	1219
bed	1242	494	465	612	609	852	1352	802	648	908	914	755	1214	462	706	855	1260	924	816	859	931	1075	876	802	934	719	577
deur	1266	825	312	641	830	829	1349	1364	671	1472	1293	528	1205	338	726	1094	1258	1470	1097	917	916	1154	1405	919	748	897	596
dicht	1412	585	646	438	566	923	1332	818	773	1180	1220	804	1133	565	813	965	1234	1115	957	739	833	1127	1155	1009	991	772	771
drie	847	541	691	829	267	676	600	815	696	1369	1034	745	591	688	1072	834	611	857	736	371	176	678	1317	491	804	391	896
één	449	715	634	965	494	164	545	788	550	950	799	541	723	559	877	391	639	647	735	659	412	504	849	351	867	418	680
fout	265	784	972	1129	698	296	421	917	604	1178	852	958	775	878	1398	563	704	420	1067	1008	573	383	1054	466	1339	586	948
gordijn	1116	710	954	1042	688	686	991	385	546	593	902	617	862	710	604	480	817	786	562	534	759	1016	617	711	925	471	628
hoger	772	489	446	706	507	402	864	425	125	688	482	439	791	345	400	382	480	555	438	667	821	735	665	442	767	258	328
interc.	1119	1136	1084	1571	1069	771	1483	762	632	55	829	868	1325	788	779	390	1120	877	650	970	1239	1341	78	639	1259	642	567
kanaal	697	649	822	1112	642	599	854	677	444	721	150	923	1084	813	850	612	610	623	538	991	950	805	723	395	1155	459	792
lager	946	629	439	915	739	533	1063	688	327	779	794	131	760	236	318	528	729	1146	584	578	719	1023	749	495	486	429	324
lamp	500	543	560	854	501	407	595	610	448	644	688	485	397	426	751	446	514	760	509	596	418	679	733	235	709	367	543
meer	1288	659	461	771	768	764	1209	821	335	962	832	295	907	138	266	748	811	1141	580	539	802	1189	959	680	352	395	315
minder	1465	771	658	1059	889	867	1305	635	307	870	839	303	997	279	76	674	878	1141	624	660	997	1282	879	884	517	563	329
open	946	1208	1059	1372	1091	703	1258	752	581	438	1086	627	1264	653	782	215	939	1072	612	866	1126	1064	469	727	1242	540	601
radio	663	958	959	1344	716	611	667	607	512	1103	665	733	699	763	880	421	161	1029	536	627	699	786	1048	408	932	344	905
stop	665	780	940	1093	751	490	783	912	587	877	704	1028	926	893	1201	717	948	335	1005	1155	836	727	794	567	1315	635	759
telef.	958	652	655	913	596	579	959	434	411	404	574	560	870	481	501	355	569	711	273	425	639	984	429	463	710	209	507
telev.	1006	647	602	665	491	580	956	481	491	676	822	485	753	438	515	418	590	910	385	193	389	866	682	535	545	252	552
twee	514	670	741	947	448	395	436	753	598	1102	838	719	535	619	1079	527	512	633	710	425	157	373	1024	340	786	317	841
uit	426	858	757	988	664	272	742	707	456	438	703	655	942	611	922	205	752	614	664	824	717	526	391	323	1159	414	586
ventil.	853	916	853	1267	820	546	1189	634	479	124	688	710	1135	629	715	313	872	716	558	780	978	1088	148	412	1069	423	525
verwar.	492	609	610	891	495	374	682	544	322	505	523	551	751	506	757	334	470	579	443	505	581	697	515	128	775	168	617
vier	1300	851	605	911	726	647	1151	958	383	940	1024	360	877	263	339	750	940	980	755	555	758	1275	891	714	369	548	265
volume	743	534	724	913	559	521	674	530	365	797	634	601	748	457	714	436	430	589	449	443	441	713	790	344	724	149	643
wekker	1264	776	773	1149	761	666	1256	575	250	642	893	269	1003	340	299	467	964	837	645	791	1021	1130	617	784	740	583	130
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tes	3	4	3	3	4	3	3	3	3	3	2	3	2	4	2	3	2	2	3	2	3	4	3	2	4	3	3

Appendix 4

Recognition performance table for the SWL-test

SWL performance

reference words in the vocabulary ↓ spoken test utterances →

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bed	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deur	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dicht	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drie	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
een	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fout	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
gordijn	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hoger	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
interc	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
kanaal	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lager	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lamp	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
meer	0	0	1	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2	0	0
minder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
radio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
telefoo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0
twee	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
uit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ventila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
verwarm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
vier	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
volume	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
wekker	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	3	4	3	3	4	3	3	3	3	3	2	3	2	4	2	3	2	2	3	2	3	4	3	2	4	3	3

Appendix 5

Distance scores tables for the SZL-test

Average distances SZL-test

| reference words in the vocabulary   spoken test utterances ——›
|
|
↓

	aan	bed	deut	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	132	1721	1087	1301	1266	645	610	1081	963	1099	696	1043	665	1265	1196	984	884	700	848	1227	1127	291	939	478	1129	1179	1008
bed	880	350	374	402	656	388	589	418	443	822	644	465	375	381	330	716	1146	683	424	595	693	918	517	543	442	650	404
deur	969	627	222	657	616	318	609	550	385	1012	659	318	404	243	323	867	1017	828	485	552	563	998	476	582	328	516	368
dicht	1214	442	508	146	747	619	840	584	684	950	897	653	608	489	556	866	1247	1129	545	590	794	1232	838	703	607	762	693
drie	569	682	400	382	418	249	420	456	363	771	478	541	389	483	488	695	837	587	351	384	353	716	538	365	414	552	542
een	450	999	424	724	610	83	258	481	458	663	492	415	292	500	515	524	430	472	285	388	305	506	504	275	415	349	471
fout	478	1665	823	1061	744	410	263	776	836	1157	931	849	426	1047	1108	913	681	597	807	822	591	388	1101	631	870	808	775
gordijn	987	859	715	863	698	485	532	153	515	539	725	480	565	593	439	478	793	596	350	418	642	1172	464	460	557	475	345
hoger	626	687	482	691	665	436	415	370	117	706	357	403	459	451	382	660	524	565	317	548	640	932	375	362	506	514	284
interc.	1012	1244	1114	1158	1268	749	803	525	909	77	653	789	846	898	527	198	777	774	340	640	1072	1112	517	553	1002	759	513
kanaal	748	893	741	1060	831	546	652	578	414	606	140	613	682	823	492	860	570	707	384	692	792	1137	237	385	764	616	611
lager	937	859	354	818	703	465	559	442	405	899	725	121	455	299	341	750	751	800	424	584	672	1172	572	553	381	465	303
lamp	580	727	325	650	508	207	305	501	368	889	554	311	209	376	440	680	757	515	399	582	453	639	518	406	346	453	465
meer	1083	709	288	688	779	454	642	513	431	1043	823	246	432	180	320	779	903	948	529	573	610	1128	540	597	338	417	354
minder	969	662	364	678	757	436	564	434	384	589	546	209	475	276	95	540	744	669	379	468	560	1075	386	560	327	372	253
open	659	1115	781	857	963	440	513	378	609	266	646	555	568	584	387	107	560	589	282	430	684	703	496	392	621	485	332
radio	816	1527	979	1317	906	636	511	625	485	881	643	556	769	944	799	804	158	907	544	786	858	1151	726	526	858	532	684
stop	520	940	668	873	829	384	323	562	606	614	444	727	378	834	608	674	790	189	502	684	580	544	532	511	698	685	490
telef.	703	720	481	634	710	277	432	318	445	347	456	315	374	368	308	306	560	528	77	323	458	906	338	261	461	344	355
telev.	939	794	441	611	536	275	537	287	429	498	577	370	480	366	269	415	685	707	282	219	370	1004	387	394	328	257	397
twee	981	1067	523	814	566	316	504	582	645	1081	788	669	528	658	674	973	874	703	622	371	79	1111	778	612	517	459	774
uit	257	1836	1049	1236	945	530	412	1050	934	1221	830	1034	619	1264	1209	984	720	722	905	1081	861	237	1150	655	1064	1088	969
ventil.	943	619	491	844	815	434	676	389	399	638	295	441	586	475	293	769	678	686	343	446	546	1234	102	356	442	336	445
verwar.	463	999	655	799	828	411	462	440	476	626	333	528	454	617	623	585	374	700	306	516	573	778	318	101	632	417	614
vier	1096	604	277	602	734	385	673	442	552	825	731	341	465	217	286	701	1032	814	487	396	448	1149	434	557	230	369	403
volume	1012	733	516	697	813	432	601	330	507	648	535	485	495	455	360	582	582	700	353	328	382	1161	334	413	490	222	459
wekker	886	744	408	728	687	400	449	312	385	487	581	226	429	322	198	426	681	566	292	452	669	952	411	544	374	437	92
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	18	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	2	3	9	3	6	3	3	6	3	3

Minimal distances SZL-test

reference words in the vocabulary ↓ spoken test utterances ——→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	57	1086	861	1218	848	640	475	905	826	940	583	925	412	989	1127	863	785	594	616	759	986	216	868	435	781	940	931
bed	644	138	177	299	373	356	521	240	334	737	593	402	147	294	259	503	1047	606	403	400	583	859	461	448	282	504	305
deur	694	394	106	423	342	283	487	302	256	926	627	301	221	185	266	674	915	724	430	338	513	856	408	531	200	443	310
dicht	1074	268	310	120	503	469	786	384	531	914	824	493	239	275	499	668	1138	1050	475	400	729	1148	831	582	434	540	604
drie	435	427	280	278	244	214	331	292	234	702	450	490	185	285	417	533	753	538	316	167	320	654	506	341	188	457	466
een	300	594	271	475	368	75	194	280	310	570	464	361	116	404	476	354	409	428	183	283	299	414	449	236	147	262	453
fout	244	1073	640	899	390	396	189	603	710	1005	851	745	258	1005	1000	688	571	574	623	630	500	338	1024	494	610	645	753
gordijn	685	750	511	558	488	436	447	100	381	474	714	318	367	587	369	361	709	524	270	289	524	1083	397	389	374	293	292
hoger	456	553	308	560	508	332	352	229	35	657	319	275	287	411	319	520	424	553	273	259	552	877	349	297	317	412	221
interc.	856	1109	855	823	1067	593	773	388	708	48	551	638	663	726	478	160	625	589	306	317	969	984	343	479	804	624	453
kanaal	508	803	530	919	530	471	552	457	335	585	120	497	478	722	445	676	459	662	359	460	694	1072	205	259	613	543	586
lager	614	626	239	461	519	429	392	257	268	822	635	60	301	185	263	574	680	721	388	488	623	1053	530	462	173	429	281
lamp	350	401	206	414	335	172	225	302	216	813	494	206	84	352	385	510	736	423	349	433	389	604	474	335	255	440	379
meer	770	477	169	334	530	442	487	285	305	974	744	208	254	97	289	604	810	874	461	461	546	993	510	520	141	325	319
minder	680	461	242	372	559	422	444	285	262	528	476	163	281	224	73	398	675	645	349	384	488	937	353	494	201	309	217
open	560	832	562	564	709	318	402	287	435	190	574	441	379	425	361	48	471	454	248	215	627	589	343	336	414	320	290
radio	474	1320	830	994	790	577	373	382	375	828	557	512	512	879	785	632	96	856	480	527	791	1059	644	399	637	483	573
stop	393	553	474	753	474	310	269	449	491	521	367	696	200	760	532	480	716	147	415	504	551	476	467	374	453	504	387
telef.	494	523	294	307	539	167	374	227	295	338	441	250	203	274	282	208	513	496	71	208	402	805	248	207	253	288	328
telev.	666	642	269	315	342	226	459	169	311	443	568	334	269	319	251	254	610	675	235	133	307	874	343	342	215	177	316
twee	670	812	351	526	265	262	436	408	470	1036	722	613	328	624	596	761	788	674	536	165	36	1010	740	573	286	379	752
uit	175	1213	914	1140	532	520	361	861	783	1045	759	884	412	1075	1096	837	583	668	716	769	730	97	1098	608	725	952	913
ventil.	649	556	297	610	632	391	603	243	296	590	217	303	280	425	270	541	553	593	313	283	450	1096	63	329	362	221	431
verwar.	295	783	468	581	652	334	370	315	320	596	298	396	258	457	550	406	298	656	180	322	550	695	260	64	410	296	593
vier	792	418	168	316	439	365	585	251	394	761	667	312	249	145	247	497	902	713	442	296	403	1013	415	418	69	276	377
volume	758	607	375	346	509	387	473	177	353	598	509	384	284	416	336	420	487	673	284	204	274	1009	295	371	363	101	431
wekker	688	538	295	457	503	361	324	199	240	405	575	149	261	274	131	302	604	555	246	300	616	822	375	433	214	349	70
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	18	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	2	3	9	3	6	3	3	6	3	3

Maximal distances SZL-test

reference words in the vocabulary ↓ spoken test utterances ——→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	373	2213	1303	1395	1918	652	714	1237	1206	1239	860	1187	1008	1544	1333	1170	1023	807	972	1637	1200	397	1064	513	1306	1381	1118
bed	997	535	921	608	1182	435	674	673	585	878	677	511	878	521	401	878	1238	760	440	925	779	985	595	627	771	908	462
deur	1112	876	596	898	1160	369	695	804	579	1085	694	330	757	323	382	1097	1132	932	570	751	663	1100	561	669	618	615	398
dicht	1428	542	1132	242	1185	717	933	790	875	1000	939	735	1224	776	589	1088	1320	1208	637	855	871	1345	853	771	996	1049	800
drie	655	906	678	515	752	294	506	590	499	846	526	605	874	690	584	936	868	636	373	644	379	815	580	397	632	611	601
een	564	1296	686	941	1121	89	350	610	637	763	509	448	602	610	587	741	467	517	405	552	313	557	601	346	685	413	484
fout	673	2107	1081	1238	1341	418	400	899	1047	1263	1007	1015	694	1094	1211	1033	780	620	994	1081	646	488	1181	798	1117	919	818
gordijn	1131	944	1253	1259	928	565	590	198	633	589	746	667	957	599	550	622	901	668	393	663	754	1288	529	560	906	693	425
hoger	748	858	793	883	860	510	515	555	189	781	385	518	904	499	415	834	605	578	387	893	741	974	415	447	717	632	338
interc.	1227	1313	1828	1600	1398	884	858	712	1112	101	802	974	1175	1050	582	222	865	959	376	848	1151	1282	699	629	1228	924	623
kanaal	896	1037	1022	1349	1044	593	806	782	494	618	171	685	1129	908	533	1100	676	753	397	996	957	1241	287	529	954	712	629
lager	1129	1005	639	1102	1060	509	658	585	639	958	771	161	733	469	401	926	831	880	497	686	767	1248	601	658	694	522	318
lamp	757	1032	404	852	932	244	371	699	550	949	602	411	447	406	495	850	805	607	452	755	532	682	554	510	481	468	518
meer	1273	971	681	917	1205	470	816	709	611	1128	899	298	728	333	340	973	1031	1022	631	685	711	1246	598	728	625	535	384
minder	1089	873	800	947	975	462	640	559	481	697	584	266	782	354	112	727	905	693	412	614	642	1174	427	617	486	466	303
open	795	1343	1359	1168	1189	536	669	588	752	358	776	670	930	739	433	193	646	725	320	639	755	854	684	436	795	619	405
radio	1013	1723	1167	1644	992	703	638	986	608	957	763	615	1136	985	816	1003	232	958	657	1041	978	1212	810	663	1105	560	768
stop	618	1234	1058	1103	1304	422	383	674	726	713	508	771	504	902	671	776	850	232	588	963	611	583	588	587	954	901	571
telef.	828	839	919	942	981	380	480	474	580	361	465	374	768	499	349	434	610	560	81	438	539	1001	456	292	753	381	374
telev.	1076	928	881	931	768	320	588	409	529	564	590	442	883	410	294	617	808	739	313	385	433	1136	472	465	549	345	452
twee	1192	1312	759	1057	910	377	581	708	886	1145	845	753	701	696	730	1205	989	733	742	647	155	1250	799	674	865	611	796
uit	346	2394	1315	1349	1504	550	491	1321	1119	1371	938	1120	816	1415	1362	1161	797	776	1041	1378	960	325	1235	694	1337	1202	1052
ventil.	1077	738	1001	1161	1054	473	747	650	458	708	344	570	1038	573	316	1028	867	779	395	647	669	1375	148	404	636	550	471
verwar.	577	1229	941	1066	1051	477	580	578	557	652	361	602	851	804	689	785	444	745	373	690	609	855	395	154	765	544	630
vier	1225	755	769	880	1089	398	815	600	735	920	799	378	767	264	339	929	1200	916	540	469	486	1260	469	672	554	526	430
volume	1159	923	1054	1004	1070	489	831	537	590	714	575	610	734	486	386	760	704	728	442	506	481	1287	355	472	770	389	490
wekker	984	873	871	1036	977	478	519	428	542	562	594	326	656	393	236	545	794	577	339	644	717	1058	466	634	628	593	116
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	18	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	2	3	9	3	6	3	3	6	3	3

Appendix 6

Recognition performance table for the SZL-test

SZL performance

| reference words in the vocabulary

↓

spoken test utterances ——→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
bed	0	3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deur	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dicht	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drie	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
een	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
fout	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gordijn	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hoger	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
interc	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kanaal	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lager	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
lamp	0	0	2	0	1	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0
meer	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
minder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
radio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
stop	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0
telefoo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	1	1	0
twee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0
uit	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
ventila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
verwarm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
vier	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0
volume	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
wekker	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	18	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	2	3	9	3	6	3	3	6	3	3

Appendix 7

Recognition performance table for the KZL-test

KZL performance

reference words in the vocabulary ↓ spoken test utterances ——→

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
bed	0	2	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deur	0	0	3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
dicht	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drie	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
een	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
fout	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
gordijn	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hoger	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
interc	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kanaal	0	0	2	0	0	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lager	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
lamp	0	0	0	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
meer	0	0	3	3	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
minder	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0
open	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
radio	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	1	0	0	0	0	0
stop	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
telefoo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
twee	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0
uit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ventila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
verwarm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
vier	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
volume	0	0	0	0	1	1	0	0	4	0	2	1	0	1	0	0	0	0	0	1	0	0	1	0	4	3	0
wekker	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	4
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	19	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	3	3	10	3	7	3	3	6	3	4

Appendix 8

Distance scores tables for the OOV-test

reference words in the vocabulary ↓   test out of vocabulary sounds ----→

	deurdig	foonhar	foonzag	foonzag	fout1	fout2	fout3	kuch1	kuch2	micstoo	neusoph	stoel1	stoel2	stoel3	stop1	stop2	stop3	tvsoap	tvsport	tvwis
aan	771	2567	3171	3091	1077	823	635	507	1480	1250	1374	1415	1419	1029	905	1043	703	1625	2696	1473
bed	479	2253	2769	2633	1015	624	570	843	1116	751	977	883	895	566	710	755	510	1006	2577	794
deur	601	2569	3225	3021	792	549	557	744	839	805	898	594	869	530	893	723	651	1078	3011	847
dicht	684	2579	3201	2956	1137	948	637	1168	1495	695	1117	937	892	782	1098	948	952	991	3045	575
drie	442	2058	2763	2489	635	654	384	553	1097	681	657	694	778	387	828	622	631	805	2424	637
een	322	2117	2894	2639	315	448	237	392	1055	544	761	559	880	544	814	488	538	963	2454	758
fout	619	1453	1964	1714	439	596	483	653	1482	1015	1104	1034	1573	925	871	1027	718	1830	1677	1345
gordijn	382	2501	3090	2623	726	455	445	892	1196	490	1140	708	662	518	736	615	499	785	2666	742
hoger	331	2237	2958	2554	674	481	375	515	879	700	837	617	629	296	706	295	574	826	2781	786
interc	413	3206	3824	3510	875	517	493	953	1264	385	1643	1082	678	888	778	731	732	793	3191	723
kanaal	374	3289	3632	3537	667	475	538	410	715	571	850	593	535	304	827	270	648	643	3267	654
lager	573	2698	3349	2983	704	506	584	743	630	619	906	297	738	445	918	623	718	1057	3148	702
lamp	515	2007	2664	2442	569	430	539	469	688	800	700	544	991	471	644	731	536	1117	2218	842
meer	677	2636	3409	3017	789	567	578	800	862	843	1100	518	787	623	820	749	727	1106	3220	836
minder	530	2750	3616	3150	720	473	550	692	729	517	932	485	586	459	721	587	617	803	3269	509
open	345	2465	3274	2846	655	489	283	642	1237	478	1278	918	676	696	717	618	604	858	2725	776
radio	525	2983	3640	3236	479	644	589	477	964	598	886	504	927	496	1316	370	1087	1155	3166	960
stop	428	1668	2320	2069	657	348	439	689	1236	853	1123	1080	1052	758	289	811	215	1145	2042	914
telefoo	285	2502	3222	2907	471	435	301	595	861	405	953	454	559	485	609	453	519	602	2753	447
tv	446	2730	3486	3142	485	419	347	705	1026	409	980	516	560	470	874	606	618	634	2987	509
twee	607	2017	2812	2386	291	767	562	755	1309	839	883	516	998	611	922	667	780	1082	2587	878
uit	671	1908	2478	2333	734	867	645	390	1362	1084	884	1239	1614	906	1119	985	987	1941	2082	1524
ventila	437	3234	3785	3523	697	423	416	586	899	597	972	526	313	391	643	369	441	417	3503	494
verwarm	350	2854	3576	3290	491	427	255	307	906	615	1008	520	437	458	579	361	564	635	3202	636
vier	640	2616	3482	3076	776	520	515	847	987	732	1075	569	572	582	741	799	581	730	3189	608
volume	408	2575	3487	3037	431	391	316	695	1096	636	1208	530	451	604	477	493	466	692	3324	569
wekker	369	2336	2998	2617	687	331	398	736	779	442	921	580	656	447	706	602	525	897	2706	594
	deurdig	foonhar	foonzag	foonzag	fout1	fout2	fout3	kuch1	kuch2	micstoo	neusoph	stoel1	stoel2	stoel3	stop1	stop2	stop3	tvsoap	tvsport	tvwis
#tests:	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Declaration of the OOV-sounds

deurdig:	slamming of a door
foonhar:	phone loudly ringing
foonzag:	phone ringing from far
fout1, fout2, fout3:	shouting "fout"
kuch1, kuch2:	cough
micstoo:	bump of the microphone
neusoph:	nose sniffing
stoel1, stoel2, stoel3:	chair pushed aside
stop1, stop2, stop3:	shouting "stop"
tvsoap:	sounds of soap program on tv
tvsport:	sounds of sport program on tv
tvwis:	sounds of a commercial program on tv

Appendix 9

Recognition performance table for the SWL-test using context

SWL performance using context

```
| reference words in the vocabulary                spoken test utterances ---- >
|
|
↓
        | aan   | bed   | deur  | dicht | drie  | één   | fout  | gordi | hoger | inter | kanaa | lager | lamp  | meer  | minde | open  | radio | stop  | telef | telev | twee  | uit   | venti | verwa | vier  | volum | wekke |
--------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
aan     |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
bed     |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
deur    |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
dicht   |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
drie    |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |
een     |     0 |     0 |     0 |     0 |     1 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
fout    |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     4 |     0 |     0 |     0 |     0 |     0 |
gordijn |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
hoger   |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
interc  |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |
kanaal  |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
lager   |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
lamp    |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
meer    |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     4 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
minder  |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
open    |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
radio   |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
stop    |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
telefoo |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
tv      |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     1 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     1 |     2 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
twee    |     0 |     1 |     0 |     0 |     0 |     0 |     1 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     1 |     0 |     0 |     0 |     0 |     0 |     0 |
uit     |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
ventila |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |
verwarm |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     2 |     0 |     0 |     0 |
vier    |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     4 |     0 |     0 |
volume  |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |     0 |
wekker  |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     0 |     3 |
--------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
        | aan   | bed   | deur  | dicht | drie  | één   | fout  | gordi | hoger | inter | kanaa | lager | lamp  | meer  | minde | open  | radio | stop  | telef | telev | twee  | uit   | venti | verwa | vier  | volum | wekke |
#tests: |   3   |   4   |   3   |   3   |   4   |   3   |   3   |   3   |   3   |   3   |   2   |   3   |   2   |   4   |   2   |   3   |   2   |   2   |   3   |   2   |   3   |   4   |   3   |   2   |   4   |   3   |   3   |
```

Appendix 10

Recognition performance table for the SZL-test using context

SZL performance using context

| reference words in the vocabulary   spoken test utterances --->

|
|
↓

	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
bed	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deur	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dicht	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drie	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
een	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
fout	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gordijn	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hoger	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
interc	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kanaal	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lager	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lamp	0	0	2	0	0	0	1	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
meer	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
minder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
radio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
telefoo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
twee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
uit	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
ventila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
verwarm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
vier	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
volume	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
wekker	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	18	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	2	3	9	3	6	3	3	6	3	3

Appendix 11

Recognition performance table for the KZL-test using context

KZL performance using context

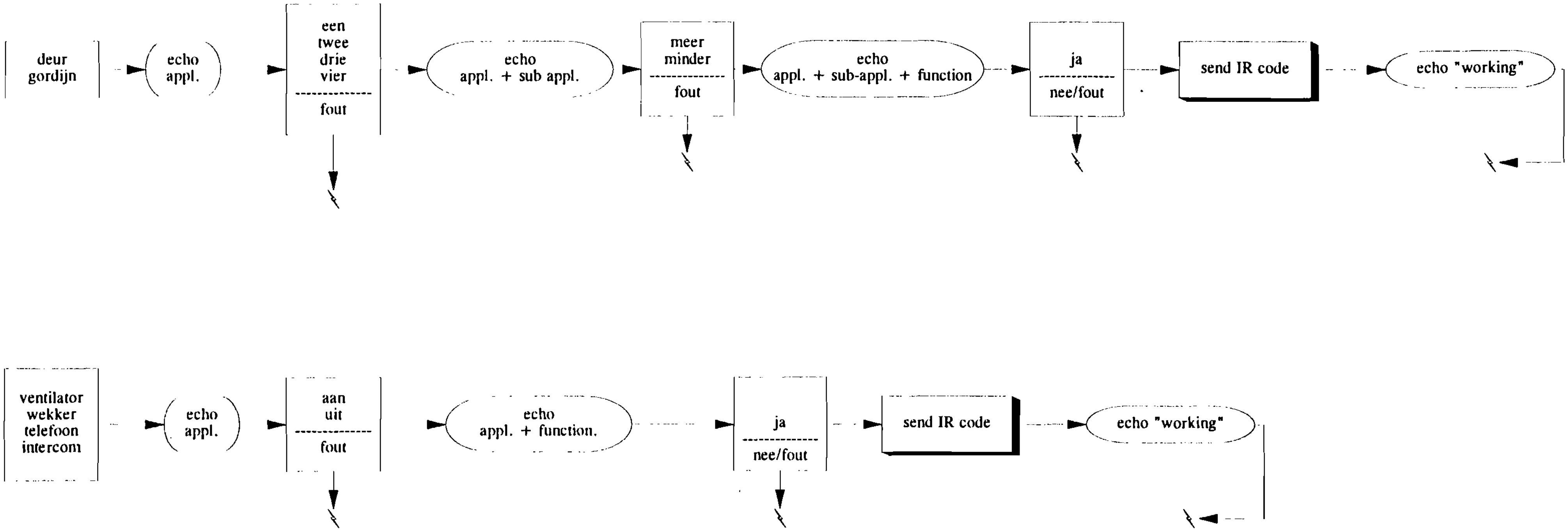
reference words in the vocabulary ↓ spoken test utterances ⟶

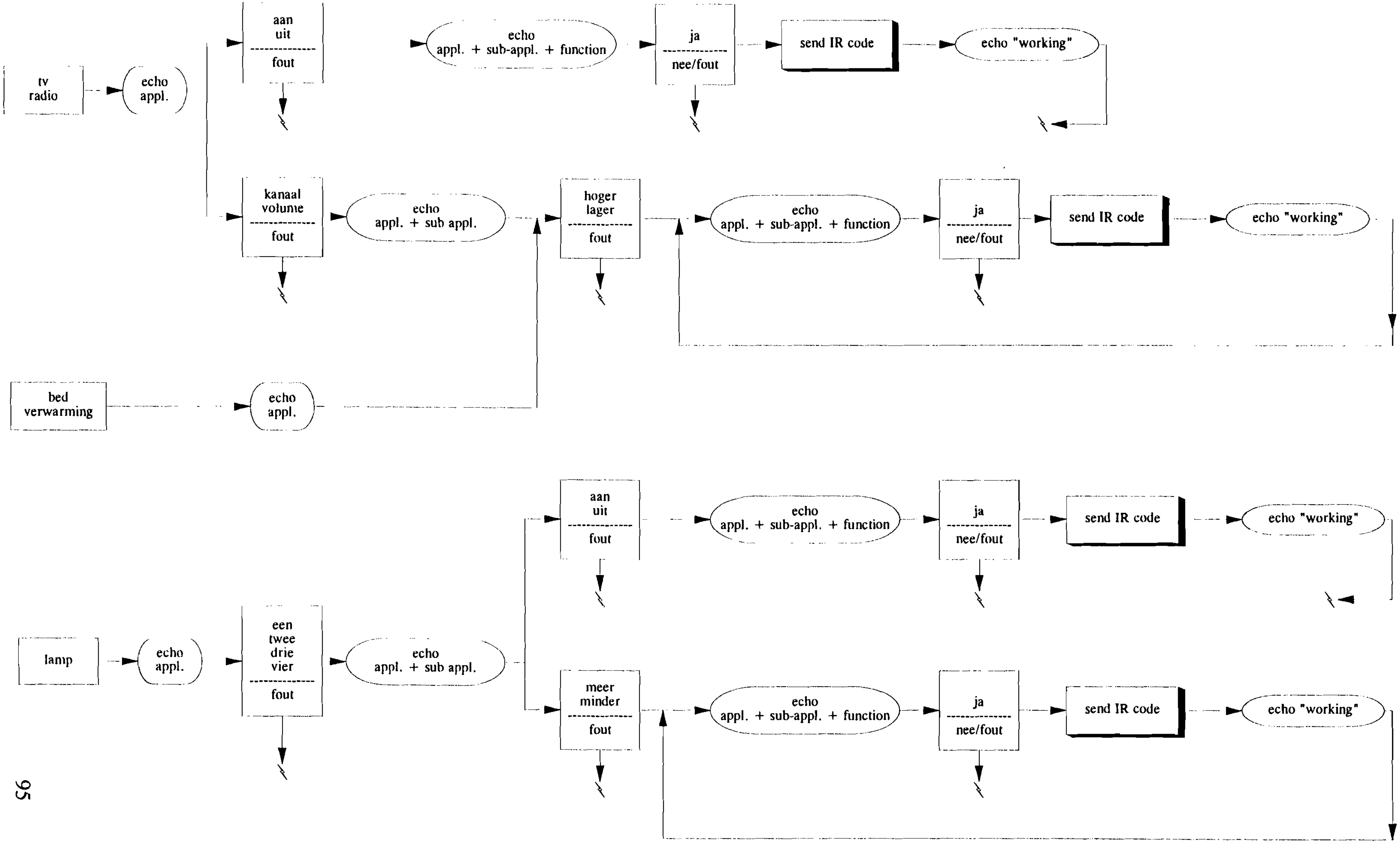
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
aan	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
bed	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
deur	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dicht	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
drie	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
een	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0
fout	7	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
gordijn	0	0	1	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
hoger	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
interc	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
kanaal	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lager	0	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lamp	0	0	4	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0
meer	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0
minder	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0
open	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
radio	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
stop	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
telefoo	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	5	0	0	0	0	0	0	0
tv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
twee	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
uit	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
ventila	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
verwarm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
vier	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
volume	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0
wekker	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	aan	bed	deur	dicht	drie	één	fout	gordi	hoger	inter	kanaa	lager	lamp	meer	minde	open	radio	stop	telef	telev	twee	uit	venti	verwa	vier	volum	wekke
#tests:	19	3	9	9	6	3	3	6	6	3	3	3	9	3	3	6	6	3	3	10	3	7	3	3	6	3	4

Appendix 12

Command dialogs

↯ = root of the command tree




tv
radio
echo
appl.
aan
uit
fout
echo
appl. + sub-appl. + function
ja
nee/fout
send IR code
echo "working"
kanaal
volume
fout
echo
appl. + sub appl.
hoger
lager
fout
echo
appl. + sub-appl. + function
ja
nee/fout
send IR code
echo "working"
bed
verwarming
echo
appl.
aan
uit
fout
echo
appl. + sub-appl. + function
ja
nee/fout
send IR code
echo "working"
lamp
echo
appl.
een
twee
drie
vier
fout
echo
appl. + sub appl.
meer
minder
fout
echo
appl. + sub-appl. + function
ja
nee/fout
send IR code
echo "working"

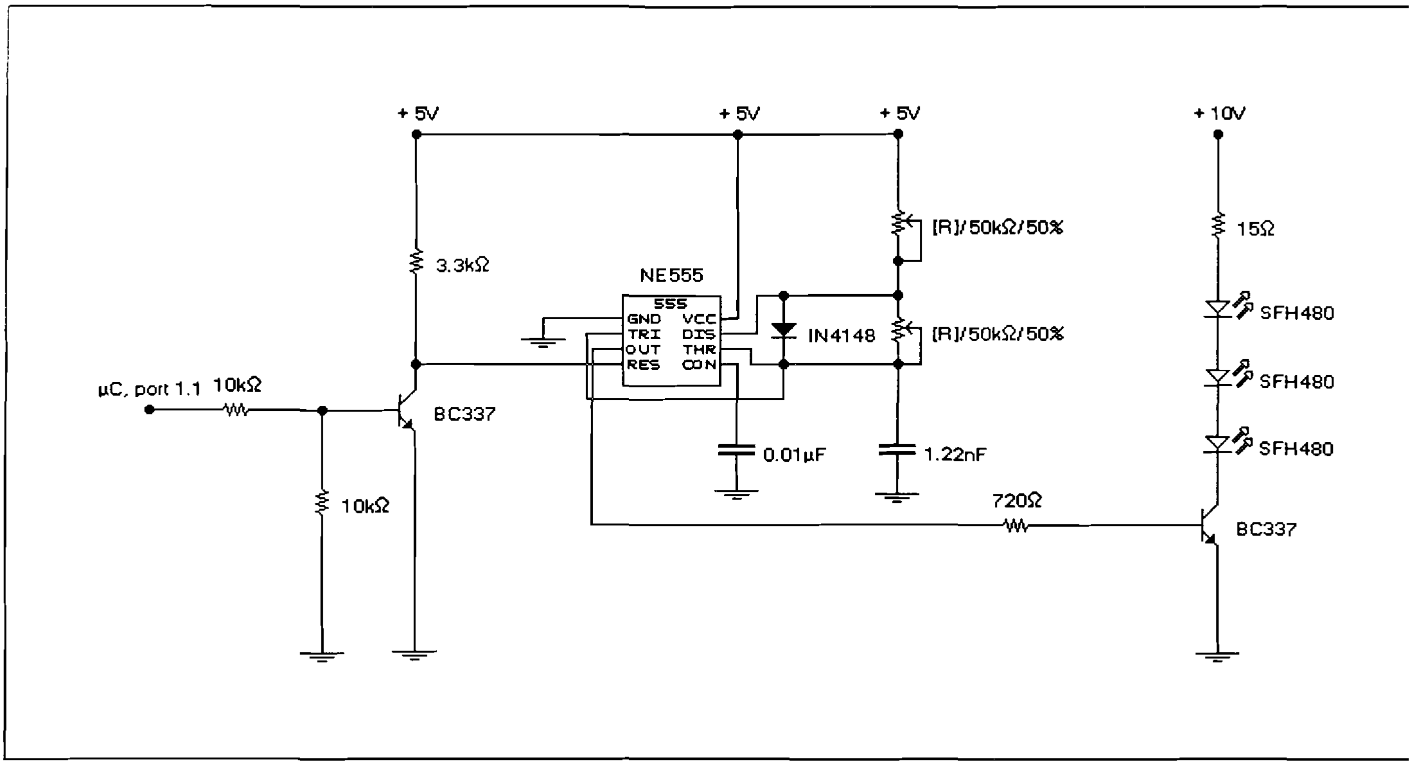
Appendix 13

Infra red codes for X-10 appliances

X-10	data bits				
	K4	K3	K2	K1	K0
device 1	1	0	0	1	1
device 2	0	0	0	1	1
device 3	1	1	0	1	1
device 4	0	1	0	1	1
device 5	1	1	1	0	1
device 6	0	1	1	0	1
device 7	1	0	1	0	1
device 8	0	0	1	0	1
device 9	1	0	0	0	1
device 10	0	0	0	0	1
device 11	1	1	0	0	1
device 12	0	1	0	0	1
device 13	1	1	1	1	1
device 14	0	1	1	1	1
device 15	1	0	1	1	1
device 16	0	0	1	1	1
function: all units on	1	1	1	0	0
function: all units off	1	1	1	1	0
function: on	1	1	0	1	0
function: off	1	1	0	0	0
function: dim	1	0	1	1	0
function: brighten	1	0	1	0	0

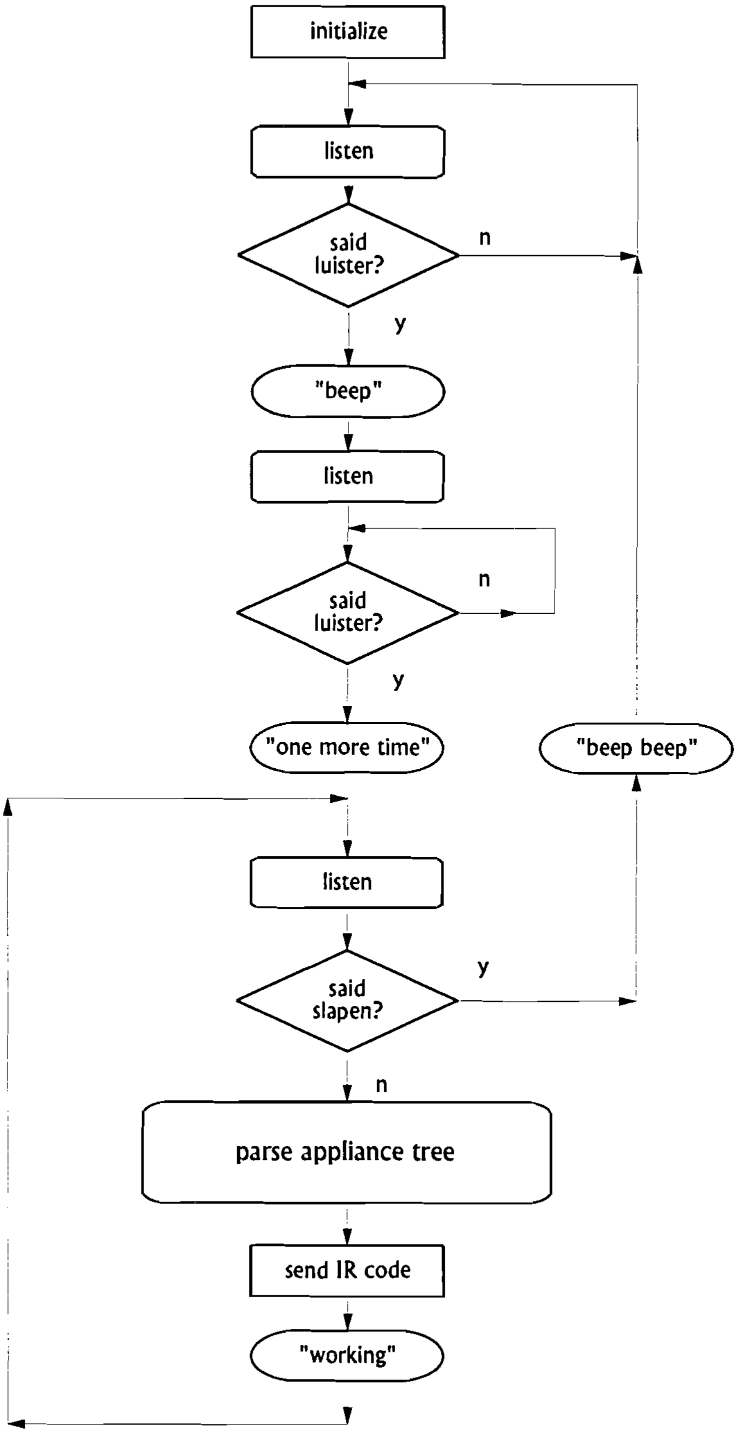
Appendix 14

Electric circuit of the infra red transmitter


+ 5V
+ 5V
+ 5V
+ 10V
3.3kΩ
NE555
555
GND
VCC
TRI
DIS
OUT
THR
RES
CON
[R]/50kΩ/50%
[R]/50kΩ/50%
IN4148
15Ω
SFH480
SFH480
SFH480
µC, port 1.1
10kΩ
BC337
10kΩ
0.01µF
1.22nF
720Ω
BC337

Appendix 15

State-machine structure of the program


initialize
listen
said luister?
n
y
"beep"
listen
said luister?
n
y
"one more time"
"beep beep"
listen
said slapen?
y
n
parse appliance tree
send IR code
"working"