Spracherkennung

Die Spracherkennungstechnologie ermöglicht die Steuerung von elektronischen Geräten durch Sprachkommandos. Sämtliche dazu erforderlichen Baugruppen sind heute bereits in moderne Schaltkreise integriert, wobei je nach Anwendungsfall eine sprecherabhängige oder eine sprecherunabhängige Erkennung erfolgen kann.

Allgemeines

Die automatische Erkennung von gesprochenen Kommandowörtern erfordert viel Know-how im Bereich der Sprachanalyse und entsprechende Rechenleistung. Die meisten Sprach-Erkennungssysteme basieren auf Softwarelösungen im PC-Bereich, da mit einem PC in der Regel genügend Rechenleistung zur Verfügung steht.

Durch Sprachkommandos mit interaktiver Menüführung kann aber gerade bei der Steuerung von elektronischen Geräten ein erheblicher Komfortgewinn erreicht werden. Dies erfordert aber hochintegrierte Spracherkennungs-ICs, die alle dazu erforderlichen Baugruppen enthalten.

Der weltweit führende Hersteller auf diesem Gebiet ist die kalifornische Firma Sensory, deren patentiertes Verfahren auf Basis eines neuronalen Netzes arbeitet. Bei den Chips von Sensory handelt es sich in erster Linie um frei programmierbare Mikrocontroller mit der zugehörigen Signalverarbeitungs-Hardware in einem Chip (Abbildung 1).

Mittlerweile steht schon die dritte Generation von Sensory-Spracherkennungs-ICs zur Verfügung. In diese hochintegrierten Bausteine sind auch die erforderlichen analogen Baugruppen, wie z. B. der Mikrofon-Vorverstärker integriert, sodass nur noch sehr wenig externe Beschaltung erforderlich ist.

Zur interaktiven Menüführung ist in diesen ICs neben der Spracherkennung auch eine Sprachausgabe enthalten.

Bei der Spracherkennungstechnologie geht es um die Erkennung von einzelnen Kommandowörtern, wobei am Anfang und am Ende jeweils eine entsprechende Pause vorhanden sein muss. Die Erkennung von einzelnen Wörtern aus einem gesprochenen Text ist nicht möglich und in den meisten Anwendungsfällen auch nicht erforderlich oder gewünscht.

Je nach Anwendungsfall kann eine sprecherabhängige (z. B. bei der Prüfung von Zutrittsberechtigungen) oder eine sprecherunabhängige Erkennung erfolgen.

Anwendungsgebiete für die Sprachsteuerung gibt es viele. So können Geräte per Sprachkommandos bedient werden, die über eine interaktive Menüsteuerung auch Rückmeldungen ausgeben, und das tastenlose Telefon ist keine Utopie mehr.

Bei der sprecherunabhängigen Erkennung hat das System einen fest eingebauten Wortschatz, der ohne vorheriges Trai-

Bild 1: Im Gegensatz zum RSC 300 ist beim RSC 364 ein 64-kB-ROM vorhanden.

ning Kommandos von unterschiedlichen Sprechern erkennt. Hierfür muss aber vor der Entwicklung genau bekannt sein, welche Wörter erkannt werden sollen und wie der Kreis der Benutzer zusammengesetzt ist (männlich, weiblich, beide, Erwachsene, Kinder, Sprache, Dialekt usw.). Anschließend müssen repräsentativ ca. 400 bis 500 Aufnahmen des Wortschatzes gemacht werden, die dann von Sensory aufbereitet werden müssen. Ein sprecherunabhängiges Set kann max. 14 Wörter enthalten, wobei kleinere Sets bessere Ergebnisse ermöglichen.

Bei der sprecherabhängigen Erkennung berücksichtigt das System die individuellen Sprachmerkmale eines Sprechers anhand eines vorher trainierten Schlüsselwortes. Das System ist somit auch zur Zutrittskontrolle (biometrische Sicherheit) geeignet.

Betrachten wir nun die sprecherabhängige Spracherkennung näher. Hier muss das Kommandowort oder auch mehrere Wörter, wenn verschiedene Schaltfunktionen erfolgen sollen, vom System (Spracherkennungs-Chip) erst trainiert werden.

Beim Voice-Direct-364-Sprachmodul von Sensory z. B. können vom System bis zu 15 unterschiedliche sprecherabhängige Kommandowörter trainiert werden. Jedes trainierte Wort oder Phrase muss dabei kürzer als 2,5 s sein und darf keine Pausen, die länger als 0,5 s sind, enthalten.

Es ist jedoch nicht grundsätzlich erforderlich, alle 15 Wörter in einem Trainingsdurchlauf anzulegen. Das Training kann jederzeit nach einer beliebigen Anzahl von Wörtern beendet werden. Bei Bedarf kann zur Vervollständigung des Sets das Training zu einem späteren Zeitpunkt wieder aufgenommen werden.

Das Löschen von einzelnen Sprachmustern (Wörter oder Phrasen) ist beim Voice-Direct-364-System nicht möglich. Hier ist nur das komplette Set zu löschen und durch ein neues zu ersetzen.

Einen erheblichen Einfluss auf die Qualität der Spracherkennung haben bei jedem System die Umgebungsbedingungen, wie z. B. Hintergrundgeräusche, während der Trainingsphase und auch während der späteren Spracherkennung. Die besten Ergebnisse werden erreicht, wenn die Trainingsaufnahme in der Umgebung durchgeführt wird, wo später das Endprodukt eingesetzt werden soll.

Einen wesentlichen Einfluss auf die Erkennungsqualität hat die Art der Hintergrundgeräusche, wobei natürlich das Sprachsignal sich immer deutlich vom Hintergrundsignal abheben muss.

Ungleichförmige Nebengeräusche (Radio, TV) sind wesentlich problematischer als gleichförmige Geräusche, die z. B. von einem Lüfter stammen.

Es bringt keinen Vorteil, das Training in einer ruhigen Umgebung durchzuführen, wenn das Produkt später in einer Umgebung mit Hintergrundgeräuschen zum Einsatz kommt.

Neben der Geräuschkulisse gibt es einige weitere Faktoren, die grundsätzlich auf die Qualität eines Spracherkennungssystems einen wesentlichen Einfluss haben. So sollte die Entfernung zwischen dem Mund und dem Mikrofon während der Trainingsphase ungefähr die gleiche sein, wie beim späteren Einsatz.

Kurze Distanzen sind grundsätzlich vorteilhaft, da bei größerer Entfernung auch die Raumakustik (Echo, Hall) einen erheblichen Einfluss hat.

Auch hat es einen nachteiligen Einfluss, wenn unterschiedliche Mikrofone oder Gehäuse im Mikrofonbereich zum Einsatz kommen.

Die Aussprache der Kommandowörter und die Stimmlage während der Trainingsphase sollte der normalen Sprachwiedergabe entsprechen. Unnatürliche Betonungen oder Akzente verursachen eine schlechte Wiedererkennung. Während der Trainingsphase sollte die körperliche und emotionale Verfassung nicht wesentlich vom typischen Einsatzfall abweichen.

Durch eine sorgfältige Auswahl der Kommandowörter kann eine erhebliche Steigerung der Erkennungsgenauigkeit erreicht werden. Gleichklingende Wörter oder Phrasen, wie z. B. „aus" und „Haus" begünstigen natürlich Erkennungsfehler. Die Auswahl von Wörtern mit unterschiedlicher Silbenzahl dagegen erhöht die Erkennungsgenauigkeit.

Sehr gute Ergebnisse werden mit der sprecheradaptiven Spracherkennung erreicht. Bei dieser Methode wird bei der ersten Benutzung eines Systems ein vorprogrammiertes, sprecherunabhängiges Set benutzt. Bei jeder weiteren erfolgreichen Erkennung wird die individuelle Aussprache mit dem vortrainierten Wort gemittelt, sodass sich der Algorithmus an die Sprechweise des Benutzers anpasst. Die Erkennungsgenauigkeit wird dadurch im Laufe der Zeit immer besser. Zum Einsatz dieser Technologie muss aber auch eine Korrekturmöglichkeit vorgesehen werden, um dem System mitzuteilen, wenn ein Wort falsch erkannt wurde.

Eine weitere Möglichkeit ist die sogenannte Dual-Recognition-Technologie. Hier wird zunächst ebenfalls von einem sprecherunabhängigen Wortset ausgegangen, und dann die individuelle Aussprache des Benutzers einmalig aufgezeichnet und mit dem vorprogrammierten Wort gemittelt. Die Erkennungsgenauigkeit ist höher als bei der sprecherunabhängigen Erkennung, der Algorithmus passt sich aber nicht kontinuierlich an. Diese Methode ist am besten zu nutzen, wenn keine Korrekturmöglichkeit vorgesehen ist.

Für die Erkennung von Ziffern zur Eingabe numerischer Werte gibt es einen optimierten Algorithmus unter der Bezeichnung „fast digits".

Wenn das System nicht durch eine Tastenbetätigung aktiviert werden soll, besteht die Möglichkeit, ständig nach einem bestimmten Schlüsselwort zu hören (Continuous listening). Bei der Erkennung wird dann automatisch die gewünschte Aktion ausgeführt. Diese Technologie funktioniert sowohl sprecherabhängig als auch sprecherunabhängig.

Doch nun zurück zu den Spracherkennungs-ICs von Sensory. Neben den einzelnen ICs wird von Sensory unter der Bezeichung „Voice Direct 364" auch ein komplett programmiertes Spracherkennungsmodul angeboten. Dieses Modul ist mit der maskenprogrammierten Variante des leistungsfähigsten Sensory-Spracherkennungs-Chips, dem RSC 364, ausgestattet. Das Programm des Moduls befindet sich dabei in einem chipinternen 64-kB-ROM.

Das Modul (Abbildung 2), dessen Schaltungstechnik wir in einem weiteren Artikel

vorstellen, führt durch Mustervergleich mit vorher trainierten Sprachmustern in Echtzeit eine deutschsprachige sprecherabhängige Erkennung diskreter Wörter oder Phrasen durch.

Neben der Spracherkennung verfügt dieses Modul zur interaktiven Menüführung auch über eine Sprachausgabe (Ansagetexte) in Deutsch. Das Voice-Direct-Modul kann wahlweise als Stand-alone-Anwendung oder zur Anbindung an einen externen Mikrocontroller konfiguriert werden.

Wie bei jeder sprecherabhängigen Spracherkennung müssen auch beim Voice-Direct-Modul die einzelnen Kommandowörter zuerst trainiert werden. Während des Trainings erzeugt das Voice-Direct-Modul dann Sprachmuster, die der individuellen Stimme des Sprechers entsprechen und speichert diese in einem nicht flüchtigen EEPROM ab. Das während der Erkennung neu erzeugte Sprachmuster wird mit den abgespeicherten Daten verglichen. Bei hinreichender Übereinstimmung erfolgt dann die Freischaltung des entsprechenden Ausgangs.

Folgende Aktionen führt das Voice-Direct-Modul bei jeder Erkennungssequenz durch:

1. Verstärkung und Filterung des vom Mikrofon kommenden Audio-Signals (gesprochenes Wort oder Phrase). Dieses analoge Signal wird dann in digitale Werte umgewandelt.
2. Das Modul analysiert dann den Sprachsignalverlauf und erzeugt ein Muster von Informationen, die die signifikanten Sprachmerkmale repräsentieren.
3. Um eine gute Signalqualität zu erreichen, erhöht oder verringert Voice Direct 364 automatisch den Verstärkungsfaktor.
4. Das erzeugte Muster wird mit Hilfe eines neuronalen Netzwerkes mit den vorher gespeicherten Sprachmustern verglichen. Zuerst wird eine kleine Zahl in Frage kommender Muster ausgewählt.
5. Die ausgewählten Muster werden weiterverarbeitet, um das eine Muster zu ermitteln, das am besten dem vorgegebenen Sprachmuster entspricht.
6. Wenn das ausgewählte Sprachmuster eine Übereinstimmung mit den abgespeicherten Werten aufweist, die oberhalb eines vordefinierten Grenzwertes liegt, ordnet Voice Direct 364 das empfangene Muster einem vorher trainierten Wort zu. Wenn zu keinem der vorher trainierten Wörter eine Übereinstimmung oberhalb des Grenzwertes festgestellt werden kann, erfolgt eine Ablehnung.

Die ersten drei Schritte sind für jedes Wort während des Trainings zu wiederholen. Um die Genauigkeit zu erhöhen, werden dann von jedem Wort zwei Sprachmuster erzeugt und der Durchschnitt von beiden Sprachmustern abgespeichert.

Bevor ein Sprachmuster in den Speicher übernommen wird, erfolgt ein Vergleich mit den bereits bestehenden Einträgen des Sets. Bei zu großer Ähnlichkeit mit einem abgespeicherten Sprachmuster des Sets wird das neue Wort nicht akzeptiert.

Auch wenn die Erkennungsgenauigkeit des Voice-Direct-Systems typisch bei 99 % liegt, gibt es grundsätzlich, wie bei jedem Spracherkennungssystem, auch zwei Arten von Fehlern:

1. Ein abgespeichertes Wort wird nicht erkannt und somit kein Schaltvorgang ausgelöst.
2. Ein nicht bekanntes Wort wird mit einem abgespeicherten Wort verwechselt und dadurch ein falscher nicht erlaubter Schaltvorgang ausgelöst.

Beim Voice-Direct-364-Modul kann die Erkennungsselektivität eingestellt und an die individuellen Bedürfnisse angepasst werden. Je nach Einstellung kommt es dann zu mehr Vertauschungsfehlern und dafür weniger Zurückweisungen bei korrekten Kommandowörtern oder umgekehrt. Die beste Einstellung ist abhängig von den Umgebungsbedingungen und somit am besten experimentell zu ermitteln.

Unter der Bezeichnung Voice Extreme™ steht ein weiteres, umfangreiches Toolkit von Sensory zur Verfügung (Abbildung 3). Das eigentliche Sprach-Erkennungsmodul befindet sich dabei auf einer Experimentierplatine, die über umfangreiche Anschlussmöglichkeiten verfügt. So steht z. B. eine RS-232-Schnittstelle zum Anschluss an einen PC zur Verfügung. Damit sind dann eigene sprachgesteuerte Anwendungen in einer Windows-Entwicklungsumgebung realisierbar und zum Voice-Extreme™-Modul herunterzuladen.

Für eigene Hardware-Erweiterungen steht auf der Entwicklungsplatine eine Lochrasterfläche zur Verfügung. Auf einer zum Lieferumfang gehörenden CD ist ein umfangreiches Programmpaket zusammengestellt, das einen Editor, C-Compiler, Linker, Help-Funktionen und einen Downloader enthält. Mit der Software können auch WAV-Dateien in ein für das Modul zu verarbeitendes Format konvertiert werden.

Der Aufbau und die detaillierte Schaltungstechnologie des Voice-Direct-364-Moduls wird im „ELVjournal" 3/2003 ausführlich vorgestellt.

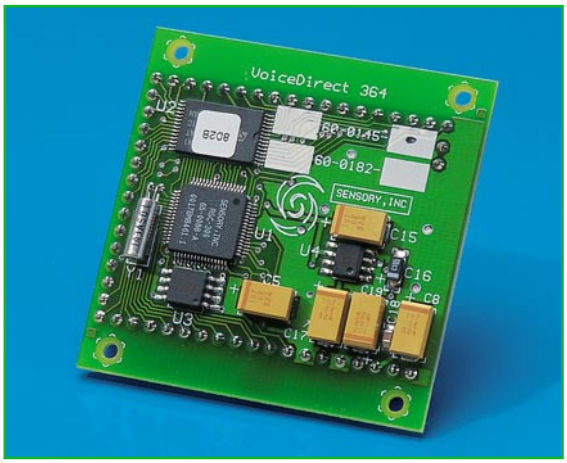

Bild 2: Komplettes Spracherkennungsmodul von Sensory mit dem RSC 364

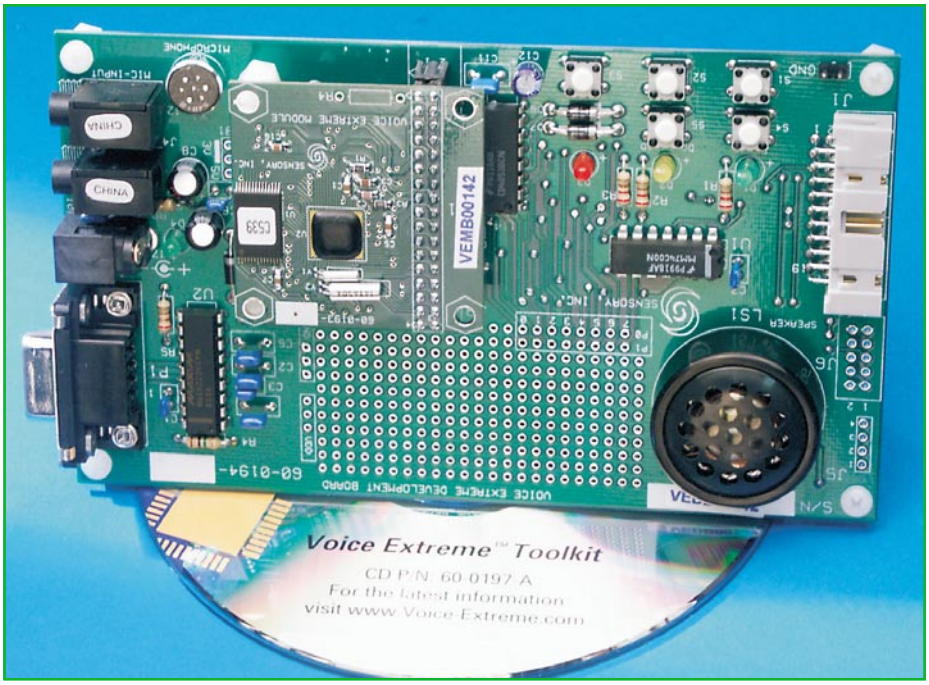

Bild 3: Das Voice Extreme™ Toolkit ermöglicht die Entwicklung von eigenen, sprachgesteuerten Anwendungen.