Autor: Christian Voit

Spracherkennungs-Grundlagen

In jedem besseren Science-Fiction-Film hat man es schon gesehen: Computer und alle Arten von technischen Gerätschaften können Sprache verstehen und reagieren auf Kommandos, meistens in Form einer regelrechten interaktiven Unterhaltung. Diese Raumschiff-Enterprise-Fiktion ist zwar heute noch keine Realität, doch die ersten Ansätze sind bereits im täglichen Leben zu beobachten: Mobiltelefone, die mit Sprachwahl funktionieren, Autos, in denen man per Wortkommando verschiedenste Aktionen steuern kann, und nicht zuletzt Diktierprogramme für den PC. Doch was steckt eigentlich an Technologie hinter diesen Systemen, und warum sind sprachgesteuerte Geräte im täglichen Leben immer noch relativ selten anzutreffen?

Allgemeines

Die Erkennung von Sprache ist ein sehr komplizierter Vorgang, denn im Gegensatz zu einem Tastendruck, der praktisch 100 % zuverlässig erkannt werden kann, wird ein Wort selbst von ein und demselben Sprecher nie zweimal exakt gleich ausgesprochen. Die vom Stimmorgan produzierten Schallwellen können sich bei jedem Sprechvorgang sogar erheblich voneinander unterscheiden. Weiterhin ist die Stimme jedes einzelnen Menschen so individuell wie sein Fingerabdruck, es gibt keine zwei exakt gleichen Stimmen. Und als ob das alleine nicht die Spracherkennung schon schwierig genug machen würde, gibt es dann noch die unzähligen Nebengeräusche und die akustischen Eigenschaften der Umgebung, denn wir befinden uns ja normalerweise nicht im reflexionsfreien und schallisolierten Tonstudio. Zudem enthalten praktisch alle Sprachen Homophone, d. h. Wörter mit gleichem Klang, aber verschiedener Bedeutung (Meer / mehr). Es gibt also gleich eine ganze Menge von Gründen, die die Spracherkennung erschweren. Aber betrachten wir doch zunächst einmal, was beim Sprechen bzw. Hören geschieht.

Spracherzeugung

Der menschliche Stimmapparat ist ein hochkomplexes Organ. Die Grundschwingung wird durch die Stimmlippen im Kehl-

kopf erzeugt, die umgangssprachlich auch oft als Stimmbänder bezeichnet werden. Mehrere Muskeln können die Masse, Länge und Spannung der Stimmlippen beeinflussen. Zusammen mit der durchströmenden Luftmenge aus der Lunge können wir damit eine sehr große Bandbreite hinsichtlich der Tonhöhe und der Lautstärke der Stimme erzeugen. Die Frequenz der Grundschwingung beim Sprechen liegt im Durchschnitt bei rund 120 Hz bei Männern und ca. 250 Hz bei Frauen. Eine Hochgeschwindigkeitsaufnahme der schwingenden Stimmlippen kann man unter http://www.iis.fraunhofer.de/medtech/med_bild/stilip/index_d.html sehen.

Die Stimmlippen erzeugen meist eine annähernd dreieckige Wellenform, somit sind neben dem Grundton auch Oberwellen enthalten. Im Resonanzraum, der von Rachen, Gaumen und Zunge gebildet wird, sind wir in der Lage, durch gezielte Verstärkung und Abschwächung von Oberschwingungen Vokale zu bilden. Ein Monophthong ist ein einfacher Vokal ohne Veränderung der Qualität (A, E, I, O, U, Ä, Ö, Ü), wohingegen Übergänge zweier Vokale als Diphthong bezeichnet werden (ei, au …).

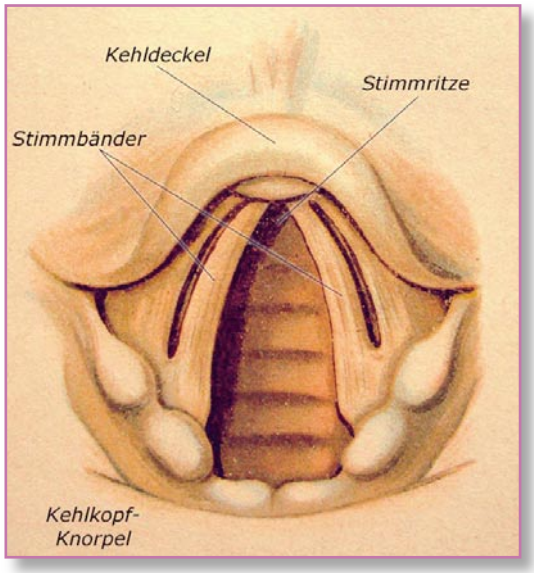
Doch neben den Vokalen, die auch als stimmhaft bezeichnet werden, gibt es noch eine Vielzahl weiterer Laute, die zur Artikulation von Sprache notwendig sind. Die stimmlosen Laute entstehen im Mund- oder Nasenraum. Grob gruppieren kann man sie in Frikative, welches Zischlaute wie S, F, oder Sch sind, die durch Luftverwirbelungen an Engstellen erzeugt werden, sowie in Plosive wie P, T oder K, die durch komplette Blockierung und plötzliches, „explosives“ Freigeben des Luftweges durch Zunge und Lippen entstehen.

In der Phonetik und Phonologie werden diese Laute als Phone bezeichnet. Im Internationalen Phonetischen Alphabet (IPA) sind 95 phonetische Zeichen als Unicode definiert, mit denen alle Laute aller menschlichen Sprachen genau beschrieben werden können. Zusätzlich gibt es noch eine Vielzahl von Betonungszeichen als Ergänzung. Ein Phon (hier übrigens nicht zu verwechseln mit der Maßeinheit für die Lautstärke) ist das kleinste Lautelement einer sprachlichen Äußerung. Demgegenüber ist ein Phonem die kleinste bedeutungsunterscheidende Einheit eines Sprachsystems. Ein Phonem kann aus mehreren Phonen bestehen. Während also ein Phon lediglich einen Laut bezeichnet, unabhängig vom Sinn eines Wortes, wird die Bedeutung eines Wortes durch die Abfolge von Phonemen charakterisiert (Beispiel: Bein und Pein).

Jede Sprache der Welt verwendet nur einen Ausschnitt von Phonen und Phonemen.

Im Deutschen gibt es z. B. nicht das englische th, was beim korrekten Aussprechen häufig zu größeren Schwierigkeiten führt. Im Englischen gibt es dagegen nicht das ch in „Acht“.

Bild 1:
Ansicht der menschlichen Stimmorgane



Hören

Die Erzeugung von Sprache ist bereits kompliziert, doch unser Hörorgan ist noch weitaus filigraner konstruiert (Abbildung 2).

Das Außenohr (Ohrmuschel und Gehörgang) erlaubt uns eine Richtungsortung von Schallquellen. Das Mittelohr (Trommelfell und die Gehörknöchelchen Hammer, Amboss und Steigbügel) dient der Transformation der akustischen Information vom Außen- zum Innenohr. Dort schließlich wird in der Gehörschnecke die akustische Information in Nervenreize umgesetzt, die dann im Gehirn analysiert werden. Vereinfacht gesagt, findet in der flüssigkeitsgefüllten Gehörschnecke eine Zerlegung in die einzelnen Frequenzbestandteile des hörbaren Spektrums (20 Hz bis 16.000 Hz) statt. Der Schall wird vom Steigbügel auf die Lymphflüssigkeit innerhalb der Schnecke übertragen und läuft dort als Wanderwelle entlang der Basilarmembran. Eine bestimmte Frequenz erzeugt ein Auslenkungsmaximum an einer bestimmten Stelle der Membran, was die dort befindlichen Sinneshärchen reizt und die Nervenimpulse erzeugt, die vom Hörnerv zum Gehirn geleitet werden. Das menschliche Gehör kann unglaubliche Leistungen vollbringen. Es ist nicht nur das Sinnesorgan mit der höchsten spektralen Auflösung innerhalb der erkennbaren Bandbreite, es hat auch einen enormen Dynamikbereich vom leisesten wahrnehmbaren Schalldruck (etwa 20 µ-Pascal, das entspricht 0 dBSPL; SPL=Sound Pressure Level) bis zur Schmerzgrenze von 120 dBSPL, was dem millionenfachen Schalldruck entspricht.

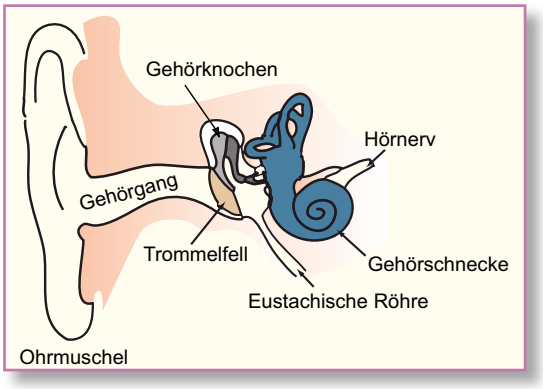


Bild 2: Die Anordnung des menschlichen Hörorgans

Technische Spracherkennung

Bei der elektronischen Spracherkennung muss zwischen verschiedenen Arten von Systemen unterschieden werden. Da gibt es zum einen die PC-gestützte Erkennung, die vorzugsweise für Diktierprogramme verwendet wird. Normalerweise kommt dabei ein Headset zum Einsatz. Zum anderen gibt es „Embedded Systems“ mit integrierter Sprachsteuerung, die im Wesentlichen zur Bedienung von Geräten per Sprachkommandos eingesetzt werden. Obwohl grundsätzlich immer ein Vergleich von zunächst unbekannten Geräuschen mit einem bekannten Vokabular erfolgt, gibt es zwischen den Diktiersystemen und der Kommandosteuerung erhebliche Unterschiede, sowohl bei den Anforderungen an

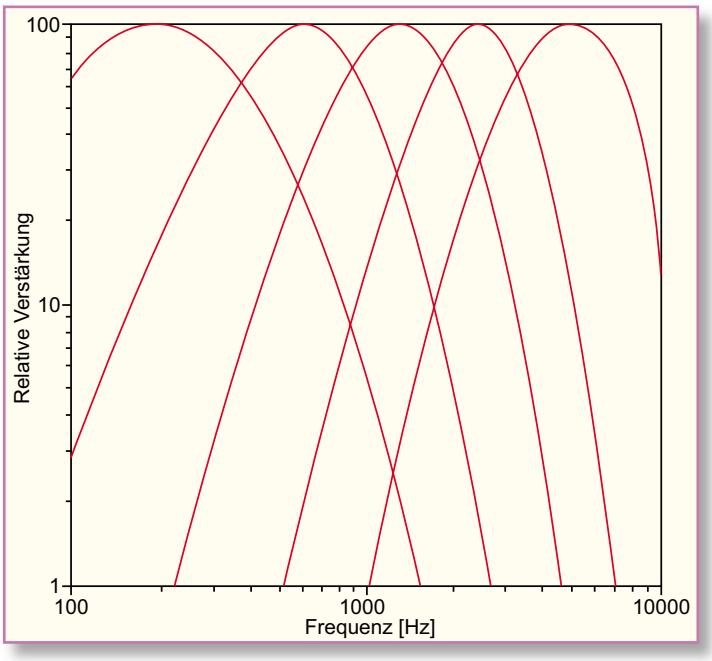


Bild 3: Filterkurve eines Spracherkenners

die Funktionsweise als auch an die dafür notwendigen Systemressourcen.

Diktiersysteme

Kennzeichnend für ein Diktiersystem sind vor allem das sehr große Vokabular und die direkte Umwandlung von Sprache in Text für die weitere Bearbeitung. Idealerweise umfasst es sämtliche Wörter einer Sprache und evtl. auch noch gängige, fremdsprachliche Begriffe.

Gute Diktierprogramme arbeiten heute mit symantischer Analyse, d. h., es wird versucht, anhand statistischer Regeln den Sinn von Sätzen zu interpretieren, um zusammen mit grammatikalischen Regeln und Modellen die Wörter in sinnvolle Zusammenhänge zu bringen und offensichtlich sinnlose Wortfolgen zu korrigieren. Die Basis dieser Diktiersysteme stellt in jedem Fall eine Phonemerkennung dar. Aus der erkannten Phonemfolge wird dann aus einem umfangreichen Wörterbuch eine Auswahl möglicher Wörter getroffen, von denen anhand von Semantik- und Grammatikregeln das wahrscheinlichste Wort ausgewählt wird.

Heutige Diktierprogramme arbeiten nur dann einigermaßen gut, wenn gewisse Randbedingungen erfüllt sind: In der Regel ist ein Training auf die Stimme des Benutzers notwendig, das aus dem Vorlesen eines dem System bekannten Textes besteht. Erkennungsfehler sollten vom Benutzer möglichst sofort korrigiert werden, weil die Programme daraus lernen können. Die Bedienung erfordert eine ruhige Umgebung und ein Mikrofon in unmittelbarer Mundnähe (Headset). Diktierprogramme erfordern leistungsfähige Rechner und viel Speicher. Sie sind nicht dafür geeignet, aus größerer Entfernung oder in lauter Umgebung bedient zu werden.

Kommandoerkenner

Im Gegensatz zu den Diktiersystemen dient ein Kommandoerkenner dazu, ein Kommandowort aus einem begrenzten Satz vorgegebener Wörter zu erkennen. Das Kommando löst dann eine bestimmte Aktion aus, z. B. das Einschalten einer Lampe oder die Wahl einer Telefonnummer. Die Sprache wird hierbei nicht in Text umgewandelt. Ziel ist vor allem die möglichst zuverlässige Erkennung von Kommandos auch in schwierigen Umgebungen, z. B. mit Hintergrundgeräuschen oder aus größerer Entfernung. Die Natur dieser Systeme setzt oft auch sehr enge Grenzen für die Kosten – sprich den Hardware-Aufwand, der für die Erkennung zur Verfügung steht. Wer würde für seine sprachgesteuerte Nachttischlampe extra einen PC aufstellen? Auch der Leistungshunger solcher embedded Spracherkenner darf mit Hinblick auf ökologische und wirtschaftliche Aspekte nur sehr gering sein, definitiv um Größenordnungen kleiner als der eines PCs.

Die Betrachtung von PC-gestützten Diktiersystemen ist sicherlich interessant, vor allem, weil sich die Technologie zurzeit noch sehr stark entwickelt und längst noch nicht ausgereizt ist. Für den Elektroniker sind aber vor allem die „kleinen" Systeme interessant, weil sich damit eine Vielzahl von Geräten, die bisher nur mit Schaltern und Reglern bedient werden konnten, in Zukunft durch Sprachkommandos steuern lassen. Und damit kommen wir dem Science-Fiction-Feeling schon ein ganzes Stück näher. Doch wie schafft es ein kleiner Mikrocontroller eigentlich, eine so komplexe Aufgabe wie die Erkennung von Sprache durchzuführen, und was steckt an Technologie dahinter? Am Beispiel des Spezial-ICs RSC4128 des Herstellers Sensory wollen wir uns die Funktion der Spracherkennung einmal im Detail anschauen.

Sprache, das ist aus akustischer Sicht der zeitliche Verlauf eines Frequenzgemisches. Die Bandbreite natürlicher Sprache liegt im Bereich von ca. 100 Hz bis 8000 Hz. Die häufig zu lesende Definition der Sprachbandbreite von 300 Hz bis 3000 Hz reicht zwar für eine einigermaßen verständliche Telefonqualität aus, aber durch diese Beschneidung vor allem der hohen Frequenzen gehen wichtige Informationen verloren, z. B. die Unterscheidbarkeit von F und S, was beim Buchstabieren am Telefon häufig zu Missverständnissen führt.

Am Anfang jeder Spracherkennung steht zunächst einmal die Umwandlung der akustischen Schwingungen in digital verarbeitbare Signale. Die Schallwellen werden von einem Mikrofon in elektrische Signale umgewandelt. Ein regelbarer Verstärker sorgt für eine Anpassung an unterschiedliche Lautstärken, so dass der nachfolgende A/D-Wandler im optimalen Bereich arbeiten kann. Dort erfolgt die Digitalisierung der Analog-Signale. Für einen guten Dynamikbereich sollte der A/D-Wandler mindestens 12 Bit Auflösung haben, besser sind 16 Bit. Die Samplerate muss so hoch gewählt werden, dass die volle Bandbreite der Sprache ausgewertet werden kann. Der RSC4128 arbeitet mit knapp 20.000 Samples pro Sekunde bei 16 Bit Auflösung, so dass die analysierbare Audio-Bandbreite bis 10 kHz reicht bei einer Dynamik von 96 dB (theoretisch).

Um nun geeignete Merkmale des Audio-Signals für die Spracherkennung zu gewinnen, muss das komplexe Frequenzgemisch in seine spektralen Bestandteile zerlegt werden. Dies erfolgt durch eine digitale Filterung des Datenstroms, und hier liegt einer der Schlüssel für den folgenden Mustervergleich: Die Erkennung soll ja möglichst zeitnah mit dem Sprechen des Kommandos erfolgen. Die Verarbeitung muss also in Echtzeit erfolgen und ist somit durch die zur Verfügung stehende Rechenleistung des Prozessors limitiert. Für eine ausreichend gute Erkennung brauchen wir keine allzu große Feinheit der spektralen Zerlegung. Eine „echte" Fourier-Transformation wäre nicht nur viel zu rechenaufwändig, sondern auch unnötig. Was wir brauchen, ist eine schnelle Extraktion von den Schlüsselmerkmalen der Sprache.

Schlüsselmerkmale

Nehmen wir an, wir würden ein Sprachsignal durch eine Filterbank von mehreren Bandpassfiltern schicken, die z. B. eine Durchlasscharakteristik wie in Abbildung 3 haben.

Wenn wir nun in ausreichend kleinen Zeitabschnitten die akustische Leistung und die dominierende Frequenz in jedem der

Tabelle 1: Beispiel eines Sprachmusters										
Zeitschlitz	1	2	3	4	5	6	7	8	9	10
Leistung Band 200 Hz	1	0	3	54	104	121	63	21	8	0
Leistung Band 600 Hz	2	0	4	55	119	134	68	9	12	0
Nulldurchg. Filt. 2	124	125	118	119	119	119	122	121	125	123
Leistung Band 1100 Hz	1	13	53	105	126	56	11	30	5	0
Nulldurchg. Filt. 3	175	185	169	142	133	156	156	166	174	167
Leistung Band 2500 Hz	35	61	68	92	110	39	0	38	41	18
Nulldurchg. Filt. 4	206	206	198	189	186	181	189	201	206	206
Leistung Band 5000 Hz	72	106	87	85	106	33	4	75	79	52
Nulldurchg. Filt. 5	215	217	211	199	197	211	223	217	217	214
Gesamtdauer	38									

Bänder ermitteln, können wir ein Muster generieren, das anschließend mit bekannten Mustern verglichen werden kann. Dies ist auch das Grundprinzip der Spracherkennung im ELV-Sprachsensor. Damit die akustischen Veränderungen bei der Phonembildung erfasst werden können, genügt es, die Leistung und die Dominanzfrequenz in Zeitintervallen von <30 ms zu erfassen. Damit steht genügend Zeit zur Verfügung, um die Merkmale auch mit begrenzter Rechenleistung zu extrahieren.

Digitale Filterung

Der RSC4128 verwendet für die Zerlegung in die Frequenzbänder eine Reihe von rekursiven Gleichungen, wie sie auch in DSPs benutzt werden. Im Unterschied zu einem „echten" DSP sind diese Gleichungen jedoch in Hardware codiert, weil dies im Gegensatz zu frei programmierbaren DSPs eine wesentlich effizientere Chipausnutzung ermöglicht.

Eine detaillierte Beschreibung der Vorgänge innerhalb des Filters würde den Rahmen dieses Beitrags sprengen, so mathematisch komplex und theoretisch sind sie. Als Ergebnis der umfangreichen Rechnungen werden in speziellen Registern die Ausgangsinformationen aller digitalen Filter für die weitere Verarbeitung zwischengespeichert.

Prozessorarchitektur

Der Mikroprozessor holt alle 50 µs die digital gefilterten Werte aus den Registern ab. Die vier Hauptaufgaben des Prozessors sind:

- Analyse der gefilterten Eingangswellenform, um in Echtzeit ein Muster zu erzeugen, das die signifikanten akustischen Informationen des Eingangssignals enthält
- Ausführung einer Mustererkennung, um festzustellen, welches Kommando aus einem bekannten Set von Wörtern erkannt wurde und wie hoch die Wahrscheinlichkeit der Erkennung ist
- Synthese von Sprachmeldungen an den Benutzer
- Ausführung jenes Anwendungsprogramms, das die Funktion des Gerätes bestimmt, in das der Prozessor eingebaut ist

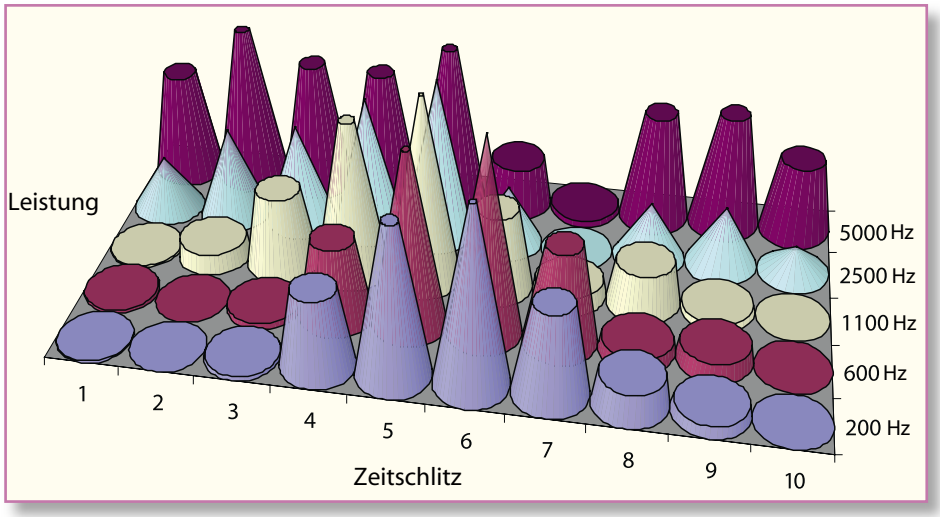


Bild 4: Die Signalleistungen der einzelnen Bänder, übertragen auf die Zeit

Blockbildung

Die niedrigen Frequenzbänder werden dazu verwendet, um die zeitlichen Grenzen des Audio-Signals zu bestimmen. Außerdem werden alle Bänder bezüglich der Signalleistung und der Anzahl der Nulldurchgänge analysiert. Dies erfolgt in zeitlichen Blöcken von ca. 25 ms. Die Anzahl der Nulldurchgänge ist ein Maß für die dominierende Grundfrequenz innerhalb eines Frequenzbandes. Als Merkmal für die Mustererzeugung berechnet der Prozessor den Logarithmus der Nulldurchgangsrate. Die Bestimmung der effektiven Leistung würde normalerweise eine Fast-Fourier-Transformation benötigen oder eine Summierung der Quadrate der Amplituden aller Datenpunkte, verbunden mit einer enormen Anzahl von Multiplikationen. Der RSC4128 verwendet ein anderes, patentiertes Verfahren, um die Signalleistung näherungsweise zu bestimmen. Dabei ist nur die Summierung der absoluten Amplitudenwerte nötig. Ein Vergleich des Logarithmus des „echten" Effektivwertes mit dem Logarithmus der absoluten Amplituden zeigt eine ausreichende Übereinstimmung bei Eingangssignalamplituden von 5 bis 160 dB. Bei der 8-Bit-Auflösung der ermittelten Werte sind die Abweichungen vernachlässigbar für die Mustererkennung.

Normalisierung

Ausgehend von den Stilleperioden, die das Sprachkommando umrahmen, nimmt der Prozessor nun eine Normalisierung der Werte vor. Dabei wird durch Interpolation bzw. Extrapolation das Geräusch auf eine fest vorgegebene Anzahl von Stützpunkten zeitlich gestreckt bzw. gestaucht. Die Lautstärke wird ebenfalls normalisiert, um unterschiedlich laute Aussprache und unterschiedliche Mikrofondistanzen aus-

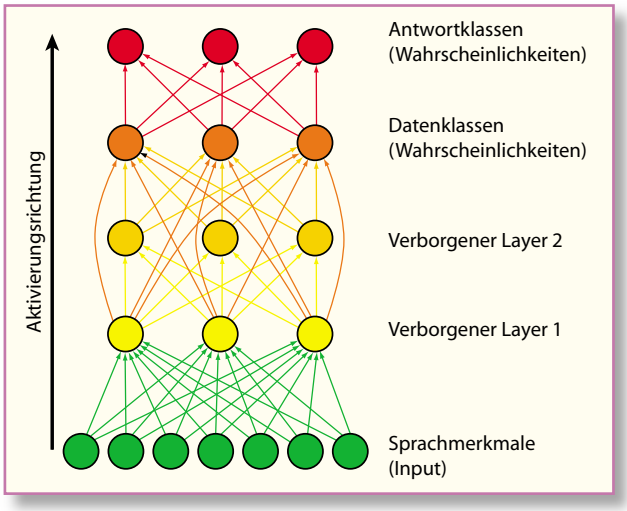


Bild 5: Die vereinfachte Struktur des neuronalen Netzes

zugleichen. Die ermittelten Dominanzfrequenzen und Signalleistungen für jeden Stützpunkt und jedes Frequenzband werden nun in Form einer Matrix als Muster zwischengespeichert und der Mustererkennung zugeführt. Tabelle 1 zeigt ein Muster, das beim Sprechen des Wortes „sechs“ erzeugt wurde.

Anschaulicher wird dieses Muster, wenn man z. B., wie in Abbildung 4, die Signalleistungen der einzelnen Bänder über die Zeit aufträgt. Es lässt sich dabei gut erkennen, wie am Anfang und am Ende des Wortes die Leistung im hohen Frequenzbereich ansteigt, während der Vokal E nur im mittleren Wortteil bei den niedrigen Frequenzen eine deutliche Leistung erzeugt.

Ein weiterer Parameter des Musters ist der Logarithmus der Gesamtdauer des Geräusches. Die Extraktion der Merkmale ist hier zur Veranschaulichung vereinfacht dargestellt. Der RSC4128 hat eine noch etwas feinere Aufgliederung der Frequenzbänder. Insgesamt besteht ein Muster aus 256 Werten. Aus der immer noch relativ geringen Zahl von Werten lässt sich leicht schließen, dass eine vollständige Rekonstruktion des Eingangssignals nicht mehr möglich ist. Man würde allenfalls einen wenig verständlichen, roboterhaften „Vokoder“-Klang erreichen. Aber die Reproduktion ist ja auch nicht das Ziel des ganzen Aufwands. Für die Spracherkennung ist die Größe der Muster völlig ausreichend und sogar vorteilhaft, denn zu viele Stützpunkte würden nicht nur die notwendige Rechenzeit für die Erkennung unnötig verlängern, dies würde auch die Erkennungsperformance verschlechtern, weil die Variationen bei wiederholter Aussprache und Nebengeräuschen zu stark ins Gewicht fallen würden.

Mustererkennung

Es gibt eine ganze Reihe von Verfahren zur Mustererkennung. Als sehr effektiv für die Spracherkennung haben sich allerdings neuronale Netze und Hidden Markov Modelling (HMM) herausgestellt. Die Besonderheit dieser Verfahren beruht darauf, dass der Mustervergleich nicht auf einer expliziten Programmierung, sondern auf einem Training des Erkennungssystems mit bekannten Mustern basiert. Zur Erkennung von Wörtern, die vom Benutzer trainiert werden können, der so genannten sprecherabhängigen Erkennung, verwendet der RSC4128 ein neuronales Netz. Eine vereinfachte Struktur des Netzes zeigt die Abbildung 5.

Die Neuronen des Eingangslayers (grün) werden mit dem bei der Sprachvorverarbeitung erzeugten Muster angeregt. Jedes Neuron eines tiefer gelegenen Layers hat Verbindungen zu jedem Neuron des darüber liegenden Layers oder sogar auch layerübergreifende Verbindungen. Jede dieser Verbindungen hat ein eigenes, so genanntes Gewicht, das positiv (stimulierend), negativ (hemmend) oder neutral sein kann. Die Summe aller Eingangsreize bewirkt in jedem Neuron eine spezifische Aktivität seiner Ausgänge, die wiederum mit anderen Neuronen vernetzt sind. Das Wissen des Netzwerkes ist in den Gewichten seiner Verbindungen gespeichert. So pflanzt sich eine Anregung der Eingangsneuronen entlang der Aktivierungsrichtung durch die verborgenen Layer bis zu den Ausgangsneuronen fort. Damit solch ein Netzwerk ein Muster zu einer bestimmten Antwortklasse zuordnen kann, muss es trainiert werden. Beim Training passiert nichts anderes, als dass die Verbindungen des Netzes neu gewichtet werden. Der eigentliche Erkennungsvorgang eines trainierten Netzes ist nunmehr kein deterministisches und transparentes Verfahren mehr, sondern unterliegt den Regeln der Wahrscheinlichkeitsrechnung. Diese Methode ist sehr effektiv, um eine tolerante Mustererkennung durchzuführen, bei der das „richtige“ Zielmuster, also ein korrekt gesprochenes Wort, durchaus Unterschiede im Detail zum trainierten Muster aufweisen darf, solange die Gesamtcharakteristik übereinstimmt.

Sprecherunabhängige Spracherkennung

Ein neuronales Netz eignet sich sehr gut für die sprecherabhängige Erkennung, d. h., wenn das Netz mit der Stimme des Benutzers trainiert werden kann, so dass das Zielmuster mit dem zu erkennenden Muster eine hohe Ähnlichkeit hat. Für eine gute sprecherunabhängige Erkennung ist jedoch noch etwas mehr Aufwand erforderlich. Statt ein komplettes Wortmuster zu erkennen, wird das neuronale Netz hierbei dazu benutzt, um eine Phonemerkennung durchzuführen. Das Netz wird dazu lediglich mit den spezifischen Phonemen für eine bestimmte Sprache trainiert, z. B. für Deutsch oder für Englisch. Der Ausgang des neuronalen Netzes liefert nun, solange gesprochen wird, einen kontinuierlichen Strom von „Phonemwahrscheinlichkeiten“ an die nächste Verarbeitungsstufe, die aus der Abfolge von Phonemen mittels einer Viterbi-Suche (eine Lösungsmethode der Hidden-Markov-Modellierung) eine Zuordnung zu bestimmten, vorgegebenen Kommandos vornimmt. Nur so ist es möglich, die Kommandos für einen bestimmten Einsatzzweck aufgrund ihrer Phonemschreibweise zu erkennen. Der Viterbi-Algorithmus ist eine Methode aus dem Gebiet der dynamischen Programmierung, die die wahrscheinlichste Abfolge von versteckten Zuständen findet, die zu einer Ab-

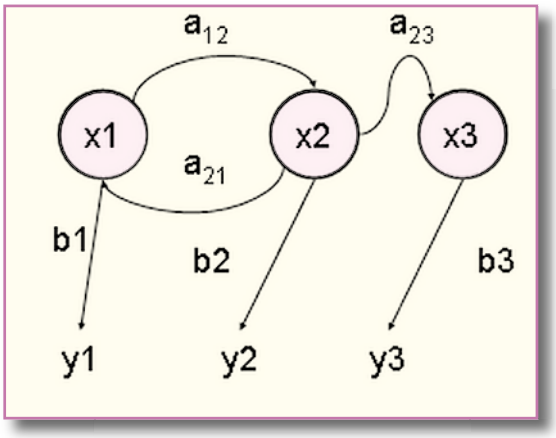


Bild 6: Das Prinzip des Hidden-Markov-Modells

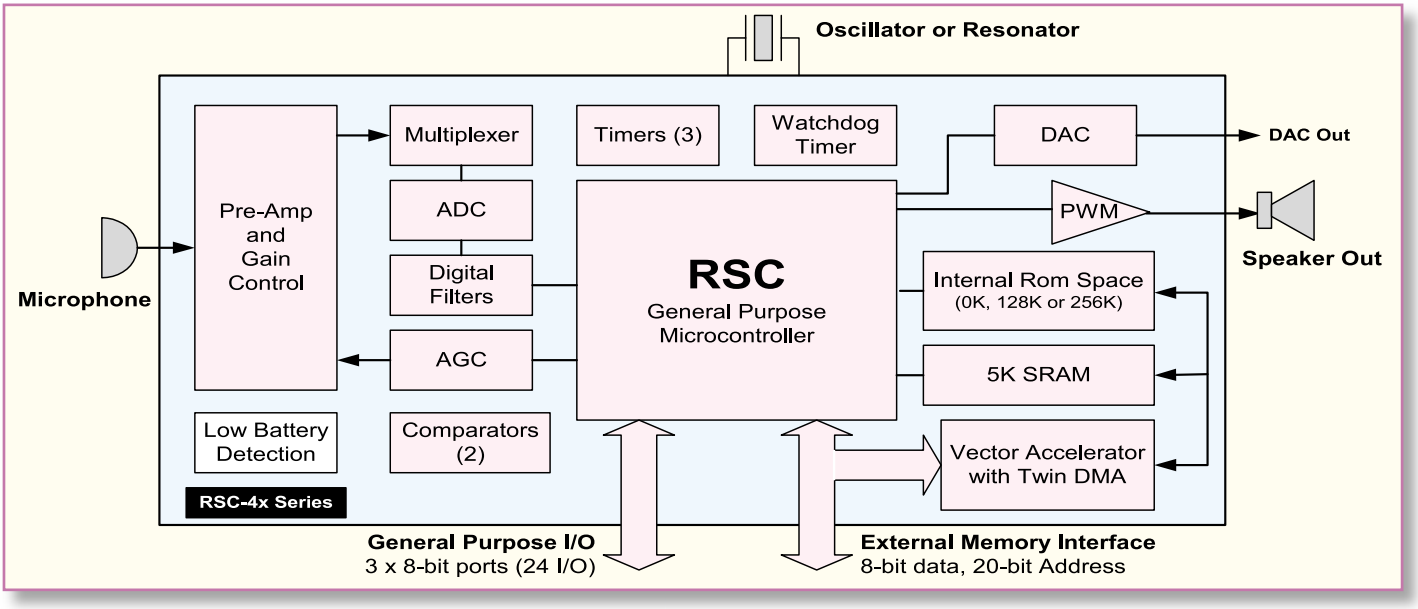


Bild 7: Das Blockdiagramm des RSC4128

folge von beobachtbaren Ereignissen führt. Er dient hier dazu, zu ermitteln, wie wahrscheinlich es ist, dass ein bestimmtes Wort einer bestimmten Folge beobachteter Laute entspricht. Abbildung 6 verdeutlicht das Prinzip des Hidden-Markov-Modells.

Es bedeuten:
- x - (verborgene) Zustände des Markov-Modells
- a - Übergangswahrscheinlichkeiten
- b - Emissionswahrscheinlichkeiten
- y - (sichtbare) Ausgabesymbole

In unserer speziellen Aufgabenstellung entsprechen die y-Werte den vom neuronalen Netz erkannten Phonemfolgen. Die x-Werte entsprechen den bekannten Phonemfolgen der vorgegebenen Kommandowörter. Durch Analyse und Vergleich der Übergangs- und der Emissionswahrscheinlichkeiten bestimmt der Viterbi-Algorithmus nun das wahrscheinlichste Kommandowort, das der beobachteten Phonemfolge entspricht.

Wir haben es also bei der sprecherunabhängigen Erkennung mit einem hybriden, zweistufigen Prozess zu tun: In der ersten Stufe erfolgt eine Vorklassifizierung von Geräuschen in Phoneme und in der zweiten Stufe wird die erkannte Phonemabfolge mit der von bekannten Kommandowörtern verglichen, um eine Übereinstimmung zu erkennen. Natürlich muss auch die Viterbi-Suche fehlertolerant sein, weil ja die Eingangsdaten, also der Phonemstrom, nur eine begrenzte Zuverlässigkeit aufweisen.

Systemarchitektur

Wie sich aus der Beschreibung der komplizierten Prozesse vielleicht schon erahnen lässt, ist ein enormer Aufwand an mathematischen Berechnungen nötig, um eine Spracherkennung durchzuführen. Ein gewöhnlicher 8-Bit-Prozessor wäre dazu allein keinesfalls in der Lage. Man braucht schon mindestens einen DSP und einen 32-Bit-Mikrocontroller, um die intensive Mathematik auf konventionellem Wege schnell genug durchzuführen. Ein solches System ist aber trotz drastischem Preisverfall auch heute noch aufwändig und zu teuer für Consumer-Anwendungen. Durch eine clevere Systemarchitektur ist es aber dennoch möglich, alle diese Aufgaben in einem kleinen, preiswerten 8-Bit-Controller zu integrieren. Abbildung 7 zeigt das Blockdiagramm des RSC4128.

Für die sehr rechenintensiven und zeitkritischen, aber sich immer wiederholenden Aufgaben, wie z. B. digitale Filterung und Vektormultiplikation, wurden spezielle Funktionsblöcke in Silizium gegossen, die dem Prozessor so viel Arbeit abnehmen, dass die restlichen Aufgaben mit Leichtigkeit von einem 8-Bit-System verarbeitet werden können. Der RSC4128 zeichnet sich dadurch aus, dass er einen Prozessorkern beinhaltet, der dem bekannten 8051 sehr ähnlich ist. Die Registerarchitektur wurde jedoch neben einigen Special-Function-Registern (SFR) für die Audio-Vorverarbeitung dahingehend erweitert, dass der Prozessor Multitasking-fähig ist, um die Mustererzeugung und die Erkennungsalgorithmen quasi parallel in Echtzeit durchführen zu können. Dies ist sehr wichtig, damit keine wesentlichen Audio-Signale während der Verarbeitung anderer Aufgaben verloren gehen.

Weiterhin beinhaltet der Chip einen Mikrofonvorverstärker, einen A/D-Wandler und einen D/A-Wandler, damit er auch in der Lage ist, Sprachausgaben für die Interaktion mit dem Benutzer zu erzeugen. Ein PWM-Ausgang kann sogar direkt zur Ansteuerung kleiner Lautsprecher benutzt werden. RAM ist als Arbeitsspeicher ebenso enthalten wie Timer, Watchdog und Powermanagement. Die Firmware kann im internen ROM untergebracht werden oder optional auch in einem externen Speicher-Baustein. Zur Kommunikation mit der Außenwelt hat der RSC4128 24 I/O-Leitungen, die frei programmierbar sind.

Sprachsteuerung für das FS20-Funk-Schaltsystem

In der nächsten Ausgabe des „ELVjournals“ wird eine Sprachsteuerung für das FS20-Funk-Schaltsystem vorgestellt, mit der es möglich ist, bis zu 4 verschiedene Geräte ganz einfach per Sprachkommando zu schalten und sogar zu dimmen. Damit wird dann wirklich ein Teil Science-Fiction zur Realität! ELV

Bildnachweise:

Bild 1: Quelle: http://upload.wikimedia.org/wikipedia/de/2/2d/Kehlkopf_beschriftet.jpg

Bild 2: Quelle: en-Wikipedia, von Benutzer Iain selbst gezeichnet, von Christian Voit nachgezeichnet und übersetzt, http://en.wikipedia.org/wiki/Image:Ear-anatomy-text-small.png

Bild 6: Quelle: http://de.wikipedia.org/wiki/hidden_markow_model