ZWECKMÄSSIGE DIMENSIONIERUNG DER VORVERARBEITUNG BEI MIKROPROZESSORGESTEUERTEN ERKENNUNGSSYSTEMEN FÜR ISOLIERTE WORTE

W. Daxer

Institut für Elektroakustik, Technische Universität München

EINLEITUNG

Einzelwort-Erkennung wird häufig mit mittleren und großen Rechenanlagen betrieben, wobei vielfach die Rechenzeit eine untergeordnete Rolle spielt. Die dabei benutzte Software ist demzufolge für Mikroprozessor-Kleinsysteme vom Prinzip her untauglich. Grund dafür ist der weniger leistungsfähige Befehlsatz üblicher 8 bit-Prozessoren (Zahlenbereich integer: 0 - 255, keine Division etc.), darüber hinaus müssen Kleinsysteme mit Rechenzeiten unter einer Sekunde (hier 250 msec) für den gesamten Erkennungsvorgang auskommen.

Diesen zunächst unvereinbaren Forderungen zu genügen, muß der Datenfluß in das Rechnersystem drastisch reduziert werden. Üblicherweise wird eine Informationsreduktion in einem analogen Vorverarbeitungssystem durchgeführt, das bevorzugt sprachrelevante Information erhält, jedoch retundante bzw. sprecherrelevante Information möglichst unterdrückt.

Die analoge Vorverarbeitung des vorliegenden Systems stellt eine Näherung des Funktionsmodells des Gehörs dar [1]. Dieses Funktionsmodell des Gehörs, ein Ergebnis der psychoakustischen Forschung bildet Übertragungseigenschaften des menschlichen Gehörs in einem elektrischen Funktionsmodell nach. Dieser Art der Vorverarbeitung liegt die Hypothese zugrunde, daß spracherzeugendes und sprachverarbeitendes System beim Menschen phylogenetisch gekoppelt entstanden sind, demzufolge Struktur und Übertragungseigenschaften des menschlichen Gehörs sich als Vorbild für Spracherkennungssysteme anbieten. In der vorliegenden Untersuchung, die in [4] ausführlicher beschrieben wird, sollte die Frage beantwortet werden, in welchem Maße die analoge Vorverarbeitung, d.h. das Funktionsmodell des Gehörs, ohne wesentliche Einbuße in der Erkennungsrate vereinfacht werden kann. Streng genommen gelten diese gewonnenen Ergebnisse nur in Bezug auf das verwendete Erkennungssystem, eine allgemeinere Gültigkeit darf aber erwartet werden.

SYSTEMBESCHREIBUNG

Im Konzept ist das System ähnlich aufgebaut wie das auf der DAGA '80 [2] vorgeführte Einplatinensystem. Die Anzahl der Analogrechenkanäle wurde jedoch von 9 auf 12 erhöht. Dementsprechend sind auch der Multiplexer und die Software geändert, so daß sich folgende Konzeption ergibt [4]: Auf ein Mikrofon mit Vorverstärker folgt eine Filterbank mit zwölf 1,5 Bark breiten aneinandergrenzenden Bandpässen. Deren Signale werden logarithmiert, gleichgerichtet und schließlich tiefpaßgefiltert. Ein Multiplexer mit nachgeschaltetem AD-Wandler liest unter Beachtung des Abtasttheorems die Abtastwerte für jeden Kanal ein. Nachdem der Anfang eines akustischen Ereignisses erkannt worden ist, werden die Abtastwerte in den Arbeitsspeicher eingelesen. Im Speicher werden diese Zahlen zu Matrizen organisiert, wobei die Zeilenfolge die Zeit, die Spaltenfolge die Frequenz repräsentiert. Dieser Vorgang wird nach Erkennung des Endes abgebrochen und die zeitliche Länge als Zeilenzahl n festgestellt.

Je nach Sprechgeschwindigkeit bzw. Wortdauer muß eine Normierung dieser Matrix

auf eine mittlere Anzahl von Zeilen erfolgen. Dies gelingt mit einer Prozedur, die Zeilen im Falle einer notwendigen Verkürzung planvoll ausläßt oder bei einer Dehnung Zeilen verdoppelt. Der Klassifikator, der diese normierte Matrix anhand eines Musterkataloges "erkennt", ist nach dem "city-block-Prinzip" ausgeführt.

MESSERGEBNISSE

Ergebnisse mit dem vollständigen System:

Erste Erfahrungen mit dem System wurden mit sprecherspezifischer Erkennung gemacht; dabei ergab sich nach 600 Einzelexperimenten (Worten) der erste Fehler. Schon um den Meßaufwand klein zu halten, wurden 8 männliche Sprecher verschiedenen Alters, verschiedenen Dialekteinschlags etc. gewählt, um einen Referenzmustersatz zu erstellen. Ohne Verringerung der wichtigen Parameter in der Vorverarbeitung (12 Kanäle, 50 dB Dynamik, 8 bit-Abtastwerte) erreicht das System sprecherunspezifisch eine Erkennungsrate von 94,7 % entsprechend einer Fehlerrate von 5,3 %. Bei allen Messungen wurde derselbe Wortschatz von 38 Worten verwendet. Änderungen bei den Parametern wurden entsprechend auch am Mustersatz vorgenommen, so daß bei diesen Messungen ein "reduziertes" System benutzt wurde.

Reduktion des Eingangsfiltersatzes:

In einer ersten Messung wurde nur jeder zweite Kanal abgetastet, somit die Kanäle 2, 4...12 null gesetzt. Diese Änderung wurde analog auch beim Mustersatz vorgenommen. Es ergab sich eine Fehlerrate von 11,5 %. Wird umgekehrt jeder geradzahlige Kanal nicht berücksichtigt, also 1, 3...11 = 0 gesetzt, wächst die Fehlerrate nur auf 8,6 %. Um einen Anschlußwert bei Telefonbandbreite gewonnener Daten [2] zu erhalten, wurden die Kanäle 1, 11 und 12 nullgesetzt, so daß die Verarbeitungsbandbreite der in [2] beschriebenen Experimente entsprach. Es ergab sich einer Fehlerrate von 9,5 %. Die Auswertung der Gründe für Fehlerkennungen zeigt, daß fehlende hochfrequente Signalanteile der Konsonanten den wesentlichen Fehlerbeitrag liefern. Um im Folgenden einen Eindruck davon zu erhalten, welche Rolle Anzahl und Bandbreite der verwendeten Eingangsfilter spielen, ohne für jeden Meßpunkt 12 neue Bandfilter realisieren zu müssen, wurde folgende Näherung gewählt: Die Aufspaltung des sprachrelevanten Frequenzbereiches von 170 Hz bis 10 kHz in ursprünglich 12 Kanäle wurde geändert, indem in jeweils 2 oder mehr aneinander grenzenden Kanälen der maximale Spannungswert unter diesen Kanälen als Wert für einen Kanal substituiert wurde, der den Frequenzbereich der betrachteten Kanäle umfassen soll. Unter dieser Voraussetzung ist Abb. 1 zu verstehen. Dort ist die Unterteilung des Frequenzbereiches in n Kanäle gegen die

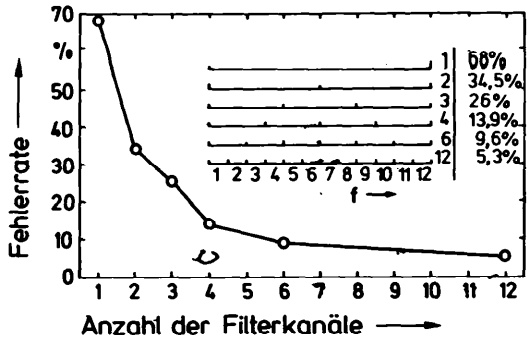

Fig.1. Fehlerrate in Abhängigkeit von der Anzahl der Bandfilterkanäle, bei festgehaltenem, erfaßten Frequenzbereich von 170 Hz bis 10 kHz.

Fehlerrate in % aufgetragen. Trotz aller Einschränkungen, die eine solche Näherung verlangt, zeigt das Ergebnis, daß eine Aufteilung in 12 Kanäle wohl eine untere Grenze sinnvoller Auflösung darstellt.

Reduktion des Dynamikbereiches:

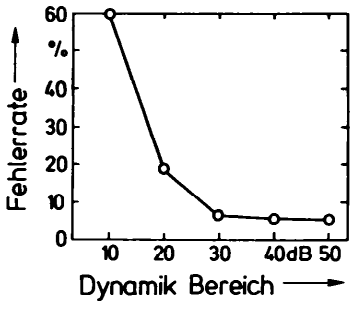

Fig.2. Fehlerrate in Abhängigkeit von dem Dynamikbereich der analogen Vorverarbeitung.

Der Einfluß des Dynamikbereiches der Vorverarbeitung auf den Erkennungserfolg ist deshalb interessant, weil diese Größe maßgeblich den Aufwand der analogen Vorverarbeitung bestimmt.

Im Experiment wurde der Dynamikbereich in 10 dB-Stufen von 50 dB bis 10 dB verringert und jeweils die Erkennungsrate gemessen. Das Ergebnis ist in Abb. 2 dargestellt und zeigt, daß die bei 30 dB Dynamikbereich erreichten Werte bei weiterer Vergrößerung nur unwesentlich verbessert werden.

Richtlinien der Amplitudenquantisierung:

Der Auflösungsgrad der Amplitudenquantisierung des Ausgangssignals der Vorverarbeitung ist deshalb wichtig, weil er direkt (als Anzahl der bits) in den Datendurchsatz eingeht und außerdem den Aufwand im AD-Wandler bestimmt. Das Ergebnis (Abb. 3) zeigt eine erstaunliche "Unempfindlichkeit" der Erkennung gegen verminderte Amplitudenquantisierung. Selbst bei einer Auflösung von 3 bit (8 Stufen!) liegt die Fehlerrate noch unter 10 %. Quantisierungsauflösungen von 10 und 12 bit (1024 bis 4096 Stufen) - wie in der Literatur oft angegeben - sind aufgrund dieses Ergebnisses nicht notwendig.

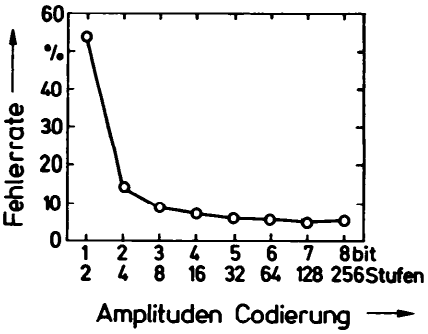

Fig.3. Fehlerrate in Abhängigkeit der Codierung der Abtastwerte in bits bzw. in Stufen.

DISKUSSION UND AUSBLICK

Es wurde die Abhängigkeit der Fehlerrate von wichtigen Parametern der Vorverarbeitung und im Falle der Amplitudenquantisierung auch des digitalen Systemes eines sprecherunabhängigen Einzelwort-Erkennungssystems gemessen. Neben der Abhängigkeit von der Zahl (pro Wort) der zur Erkennung verwendeten Muster, die

von Rabiner bereits untersucht wurde [3], sind Abhängigkeiten von der Eingangsfilter-Anordnung, von der Dynamik der analogen Vorverarbeitung und von der Amplitudenquantisierung - damit auch die notwendige Worte-Breite von AD-Wandler, Prozessor und Speicher - besonders wichtig. Unsere Ergebnisse deuten darauf hin, daß eine große Bandbreite (170 Hz bis 10 kHz) und einer Aufteilung in wenigstens 12 Filter, sowie mindestens 30 dB Dynamik zur Erzielung hoher Erkenntnisraten notwendig sind. Überraschend ist, daß bereits wenige bits in der digitalen Codierung ausreichen, um die zur Erkennung notwendige Amplitudeninformation (Pegel!) zu übertragen.

Abschließend mag gesagt werden, daß die vorliegenden Ergebnisse die Realisierung von Einzelworterkennungssystemen mit großen Wortbreiten unwirtschaftlich erscheinen lassen. Unsere Ergebnisse lassen vielmehr die Konzeption eines Feldrechensystems sinnvoll erscheinen, bestehend aus Elementen mit kleiner Wortbreite (4 bit), die relativ unabhängig voneinander im "parallel-processing"-Betrieb höhere Erkennungsleistungen bezüglich Erkennungszeit, Wortschatz und Erkennungssicherheit ermöglichen.

LITERATUR

[1] Zwicker, E.; Terhardt, E. and Paulus, Automatic speech recognition using psychoacoustic models. J.Acoust.Soc.Am. 65, 487-498 (1979).

[2] Zwicker, E. und Daxer, W., Automatische Echtzeit-Erkennung von 14 isoliert gesprochenen Worten in einem kompakten Gerät mit Mikroprozessor. In: Fortschritte der Akustik, DAGA '80, VDE-Verlag, Berlin, 731-734 (1980).

[3] Rabiner, L.R.; Levinson, S.E.; Rosenberg, A.E.; Wilpon, J.G., Speaker-independent recognition of isolated word using clustering techniques. IEEE Trans.Acoust.Speech, Signal Process ASSP-27, 336-349 (1979).

[4] Daxer, W. and Zwicker, E., On-line isolated word recognition using a microprozessor system, J.Acoust.Soc.Am.; zur Veröffentlichung eingereicht.