Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Erkennung artikulatorischer Ereignisse mit neuronalen Netzen

Guido Kolano
Gabriele Krone

Universität Ulm

Vm
Report 185
Januar 1997

Januar 1997

Guido Kolano
Gabriele Krone

Abteilung für Neuroinformatik
Fakultät für Informatik
Universität Ulm
Oberer Eselsberg
89069 Ulm

Tel.: (0731) 502 - 4242
Fax: (0731) 502 - 4156
e-mail: {guido|gabi}@neuro.informatik.uni-ulm.de

Gehört zum Antragsabschnitt: 15: Architektur

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 G gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

1 Einleitung

Im Rahmen von Verbmobil wurden *artikulatorische Ereignisse* von Julie Carson-Berndsen (Bielefeld) und Kai Hübener (Hamburg) untersucht [1, 2]. Neben diesen Arbeiten fanden auch in Ulm bei der Arbeitsgruppe von Prof. Palm (Abteilung für Neuroinformatik) Untersuchungen über die Eignung von Ereignissen als Parametersatz für die Beschreibung des Sprachsignals statt. Im Gegensatz zu den Arbeiten in Hamburg wurde sprecherunabhängig gearbeitet und es wurden keine statistischen Klassifikatoren zur Ereignisdetektion verwendet, sondern ein Ansatz mit neuronalen Netzen gewählt.

2 Artikulatorische Ereignisse

Auf dem Gebiet der Spracherkennung gibt es als erste Vorverarbeitungsstufe im wesentlichen spektrale Merkmale (Spektren, Cepstren, ...), um ein Sprachsignal zu beschreiben. Der nächste Verarbeitungsschritt führt meist schon zu Lauten (Phonen) oder Lautklassen, aus denen dann die Worte zusammengesetzt werden. In diesem Schema sind die artikulatorischen Ereignisse (im folgenden nur "Ereignisse" genannt) zwischen den spektralen Merkmalen und den Lauten einzuordnen.

In den Experimenten wurden 25 Ereignisse verwendet, die auch in Hamburg verwendet worden sind. Dabei handelt es sich um:

ap	apikal
fr	frikativ
gh	geräuschhaft
gl	glottal
la	lateral
lb	labial
na	nasal
op	Okklusionspause
pa	Pause
pl	palatal
po	palato
sh	stimmhaft
tv	Transient im Vokal
uv	
va	a-artiger Vokal
vd	dunkler Vokal
ve	velar
vg	
vh	heller Vokal
vm	mittlerer Vokal
vo	vokalisch
vr	
vt	
vu	
vz	zentraler Vokal

Leider steht kein auf Ereignisebene manuell gelabeltes Material in ausreichender Menge zu Verfügung, um damit einen neuronalen Erkenner trainieren zu können. Am IKP in Bonn (Prof. Hess) wurden zwar im Rahmen des ASL-Projektes Sprachdaten auf Ereignisebene gelabelt, jedoch ist deren Umfang zu gering. Um einen ersten Zugang zu den artikulatorischen Ereignissen zu haben, wurde auf eine Tabelle von Kai Hübener zurückgegriffen, die eine Umwandlung von Phonen in Ereignisse enthält. Im Sinne des Ereignisschemas ist dies nicht ideal, denn die Umsetzung von Phonlabeln auf Ereignislabel ist dadurch statisch, die Dynamik der Ereignisse während der Dauer eines Phons kann nicht berücksichtigt werden. Mit dieser Umsetzung von Phonen in Ereignisse kann ein möglicher Vorteil, nämlich die Ausnutzung der Dynamik auf Ereignisebene für die Phonerkennung, nicht ausgenutzt werden.

3 Ereignisvektoren als Merkmalvektoren

3.1 Idee

Eine mögliche Anwendung der artikulatorischen Ereignisse in der Spracherkennung ist die Verwendung des erzeugten Ereignisvektors als neuen Merkmalvektor. Insbesondere bei Spracherkennungsmethoden, die den zeitlichen Verlauf der Sprache berücksichtigen, ist es denkbar, daß sich durch die zusätzliche Dynamik in den Ereignissen bessere Ergebnisse erzielen lassen, da sich insbesondere die Lautübergänge mit den Ereignissen besser beschreiben lassen. Gerade die Lautübergänge sind für das (menschliche) Sprachverstehen von Bedeutung. Bei den an der Uni Ulm eingesetzten Verfahren zur Untersuchung der Lauterkennung wird der zeitliche Verlauf eines Sprachsignals aber nur insofern berücksichtigt, als daß die Eingabevektoren aus mehreren Zeitscheiben zusammengesetzt werden und so einen größeren Zeitraum umfassen. Dennoch wurde untersucht, wie sich Vektoren aus Ereignissen als Merkmalsatz zur Phonerkennung eignen.

3.2 Experimente

Um die Frage nach der Eignung von Ereignissen zu untersuchen, wurden zuerst mit einem neuronalen Netz aus konventionellen spektralen Merkmalvektoren (Spektrum, Cepstrum und PLP-Koeffizienten[1] [3]) Ereignisvektoren trainiert. Die dabei gewonnenen Ereignisvektoren wurden dann mit einem zweiten neuronalen Netz auf die Phone abgebildet. Verglichen wurden diese Ergebnisse dann mit der direkten Abbildung der Merkmalvektoren auf die Phone, wobei die dabei verwendeten Netze die gleiche Komplexität hatten wie die beiden Einzelnetze beim anderen Verfahren (Abbildung auf die Ereignisse und weitere Abbildung auf die Laute) zusammen. Erhofft wurde, daß sich durch die zweistufige Abbildung im ersten Fall die Toleranz des Systems erhöht.

Da keine echt ereignisgelabelten Daten zur Verfügung standen, wurden mittels einer Übersetzungstabelle von K.Hübener die vorhandenen Phonlabel in Kombinationen aus 25 Ereignislabeln umgesetzt. Dabei geht allerdings die zusätzliche Dynamik in den Ereignissen (gegenüber den Phonen) verloren. Mit den Lautlabeln wurden dann Sprachdaten aus dem Phondat-Korpus ereignisgelabelt (insgesamt 16 Sprecher, die "von Hand" auf Phonebene gelabelt worden waren). Ein Teil dieser Daten (von 8 Sprechern) wurde zum Training des Ereignisklassifikators eingesetzt, wobei sich für die Tests noch folgende Kombinationsmöglichkeiten ergaben:

- gleiche Sprecher, gleiche Äußerungen (Trainingsdaten)
- gleiche Sprecher, andere Äußerungen (Test_1)
- andere Sprecher, gleiche Äußerungen (Test_2)
- andere Sprecher, andere Äußerungen (Test_3)

Bei unseren Experimenten handelte es sich also um sprecherunabhängige Experimente. Es wurde bei den Tests der Datensatz Test_3 verwendet, nachdem sich gezeigt hat, daß die anderen beiden Testdatensätze nur unwesentlich bessere Ergebnisse bei den Tests zeigten. Die Kombinationsmöglichkeiten von verschiedenen Sprechern und Äußerungen sowie die verfügbare Rechenleistung des SNNS [5] auf einer Workstation schränkten die Größe der Trainings- und Testdatensätze ein.

Die für die Ereignisklassifikation verwendeten Netze hatten alle eine Schicht verdeckter Neuronen, so daß sich bei den verschiedenen Merkmalen folgende Netzarchitekturen ergaben (Zahl der verschiedenen Hidden-Neuronen in eckigen Klammern):

[1] Perceptual Linear Predictive Coding — an das menschliche Gehör angepaßtes LPC-Verfahren

1. Cepstrum (80 Koeffizienten) als Merkmalvektor (1 Zeitscheibe):
   (80 - [15, 30] - 25)

2. Fourierspektrum (160 Koeffizienten) als Merkmalvektor (1 Zeitscheibe):
   (160 - 30 - 25)

3. PLP-Koeffizienten mit Nullstellen als Merkmalvektor (1 Zeitscheibe):
   (7 - 12 - 25)

4. PLP-Koeffizienten mit Nullstellen als Merkmalvektor (3 Zeitscheiben):
   (21 - [8, 12, 24] - 25)

5. PLP-Koeffizienten mit Nullstellen als Merkmalvektor (5 Zeitscheiben):
   (35 - [8, 12, 36] - 25)

Für die Klassifikation von Phonen aus den Ereignissen wurde jeweils ein Netz der Größe (25 - 12 - 44) trainiert, da in erster Linie die Abbildung auf die Ereignisse variiert und untersucht werden sollte. Es ist nicht von vornherein klar, wie gut die verschiedenen Netze die einzelnen Ereignisse klassifizieren, deshalb wurde das Netz zur Phonklassifikation jedesmal neu auf den Ergebnissen des jeweiligen Ereignisklassifikators (mit dem Trainingsdatensatz) trainiert.

Das konkurrierende Verfahren war das Training eines komplexeren Einzelnetzes, das die gleiche Zahl von Verbindungen und eine entsprechende Topologie hat, statt der zwei getrennten Netze. Die Einzelnetze haben also die Architektur (x - y - 25 - 12 - 44). x und y entsprechen dabei den Werten der Ereignisklassifikatoren.

3.3 Ergebnisse

3.3.1 Phonerkennung

Beschränkt man sich auf Featurevektoren mit einer Zeitscheibe (Netztypen 1, 2 und 3), ergeben sich auf Phonebene keine signifikanten Unterschiede zwischen einem direkten Training und einer Vorklassifikation auf die Ereignisse mit anschließender Phonklassifikation. Dabei war es bei unseren Experimenten auch unerheblich, welche Merkmale sich in den Eingangsvektoren befanden. Ein 7-dimensionaler PLP-Vektor war ähnlich effektiv wie ein 80-dimensionaler Vektor aus Cepstralkoeffizienten oder das Ergebnis einer diskreten Fouriertransformation mit 160 Koeffizienten. Die Streuung der Phonerkennungsraten (es wurden mehrere Netze pro Architektur mit verschiedenen Initialisierungen trainiert) ist bei den Netzen, die mit Fourier- oder Cepstralkoeffizienten trainiert worden sind, größer, was auf nicht genügend Trainingsmaterial zurückzuführen ist — diese Netze sind mit so wenig Trainingsdaten nicht optimal trainierbar.

Die Experimente haben gezeigt, daß die Vorparametrisierung des Sprachsignals auf Ereignisebene weder Vor- noch Nachteile hat, was die Phonerkennungsrate angeht (Tabelle 1).

Dies ist nicht überraschend, denn die Komplexität der neuronalen Netze war in beiden Fällen gleich und der mögliche Vorteil der Ereignisse, ihre Beschreibung dynamischer Eigenschaften, wird bei der Ereignislabelung über die Phonlabelung nicht ausgenützt. Möglicherweise kann sich aber durch eine echte Ereignislabelung (nicht mehr an den Phonlabeln orientiert) ein Vorteil für die Phonerkennung ergeben. Dazu müssen aber sowohl bei der Ereigniserkennung als auch bei der anschließenden

Phon	falscher Alarm		nicht detektiert	
	direkt	kombi	direkt	kombi
.	0.0	0.0	0.0	0.0
6:	0.0	0.1	0.7	0.7
@	0.0	0.1	0.9	0.9
e:	0.0	0.0	0.0	0.0
E	0.5	6.2	1.4	0.9
9:	0.0	0.0	0.8	0.8
i:	0.0	0.0	0.0	0.0
I	8.2	1.2	2.0	3.8
Y	0.1	0.0	0.8	0.8
o:	0.0	0.0	0.0	0.0
O	0.0	1.1	2.6	2.4
u:	0.0	0.0	0.0	0.0
U	12.2	5.9	1.0	1.5
a	0.7	4.5	7.8	3.8
e:6	0.0	0.0	0.3	0.3
i:6	0.0	0.0	0.0	0.0
I6	0.0	0.0	0.5	0.5
y:6	0.0	0.0	0.0	0.0
Y6	0.0	0.0	0.0	0.0
o:6	0.0	0.0	0.0	0.0
O6	6.9	4.0	0.9	1.0
a:6	0.0	0.0	2.4	2.4
sil	0.0	0.0	0.3	0.3
k	6.7	4.6	4.7	6.6
g	0.6	0.6	3.7	3.7
N	0.0	0.0	1.1	1.1
p	0.0	0.0	0.5	0.5
b	1.0	3.0	0.4	0.4
m	0.0	0.0	0.9	0.9
R	0.5	0.9	4.5	4.2
j	0.0	0.0	0.8	0.8
v	0.0	0.0	0.1	0.1
x	0.0	0.0	0.8	0.8
S	1.1	0.9	1.3	1.5
C	1.2	0.2	0.3	0.4
f	0.9	0.8	1.9	1.6
h	1.8	2.8	2.6	2.3
t	0.1	0.0	0.6	0.6
d	5.5	7.8	2.3	2.2
n	0.0	0.0	0.8	0.8
l	5.3	6.5	3.6	2.9
z	0.0	0.0	1.1	1.1
s	0.4	0.5	0.4	0.4
U6	2.5	2.8	1.3	1.3

Tabelle 1: *Vergleich von Fehlerraten der Phonerkennung (bezogen auf alle Merkmalvektoren) bei Netzen gleicher Komplexität (1 Zeitscheibe, 7 - 12 - 25 - 12 - 44): Training PLP+Nullstellen $\rightarrow$ Phone (direkt) bzw. Training PLP+Nullstellen $\rightarrow$ Ereignisse $\rightarrow$ Phone (kombi)*

Phonerkennung die dynamischen Eigenschaften berücksichtigt werden (z.B. durch die Verwendung mehrerer Zeitscheiben oder durch rekurrente neuronale Netze).

Sowohl bei diesen Experimenten als auch bei vorherigen Experimenten mit Clusteranalyse ergaben sich für die verschiedenen Spektraldarstellungen des Sprachsignals ähnliche Ergebnisse in Bezug auf die Leistungsfähigkeit, sodaß im folgenden nur noch Experimente mit PLP-Koeffizienten einschließlich der Zahl der Nullstellen im Frame als Parameter durchgeführt wurden.

3.3.2 Ereigniserkennung

Im nächsten Schritt wurde untersucht, wie gut sich die Ereignisse an sich erkennen lassen (immer gemessen an den aus den Phonlabeln gewonnenen Labeln). Dazu wurden jetzt nur noch die PLP-Koeffizienten + Nullstellen benutzt, aber die Zahl der Zeitscheiben (1, 3, 5) und der Hidden-Neuronen (8, 12, 24, 36) war variabel. Im allgemeinen wurden wieder verschiedene Initialisierungen verwendet, die angegeben Zahlen entsprechen den Mittelwerten der Netze. Läßt man die Zahl der Hidden-Neuronen konstant und erhöht nur die Zahl der Zeitscheiben (und damit der Eingangsneuronen), ergeben sich bis auf eine Ausnahme keine systematischen Änderungen der Ereigniserkennungsraten. Die Ausnahme, bei der sich die Erkennungsrate systematisch verbessert, ist die Erkennung der Okklusionspause (op), deren Erkennungsrate von ca. 5% auf rund 25% steigt, wenn man fünf Zeitscheiben statt einer verwendet (Tabelle 2). Bei fünf Zeitscheiben ist eine bessere Abgrenzung zur Pause möglich.

An den Fehlerraten "nicht detektiert" läßt sich ablesen, daß die Erkennung von Ereignissen nicht sehr zuverlässig ist. Inwieweit dies mit der fehlenden echten Ereignislabelung zusammenhängt, läßt sich schwer abschätzen, aber viele Ereignisse treten in den Phonen, denen sie zugeordnet sind, sicherlich nur kurz auf und sollten auch nicht dem ganzen Phon zugeordnet werden. Der Vergleich mit den Ergebnissen von K.Hübener in [2] zeigt, daß bei den meisten Ereignissen ähnliche Fehlerraten auftreten, woraus man schließen kann, daß die Erkennung der verschiedenen Ereignisse unterschiedliche Schwierigkeitsgrade hat.

Die Erkennung der Ereignisse verbessert sich nicht, wenn man die Zahl der Hidden-Neuronen erhöht. In diesem Fall scheint eher eine Zunahme des Fehlers "nicht detektiert" aufzutreten, aber das läßt sich aus den Ergebnissen nicht eindeutig ableiten.

3.4 Ereignisvektoren als Merkmalvektoren — Schlußbetrachtung

Sollen Ereignisvektoren als zusätzliche Merkmale für die Phonerkennung eingesetzt werden, darf man nicht, wie hier untersucht, die Ereignislabelung starr aus der Phonlabelung ableiten. In diesem Fall entsteht zwar keine Verschlechterung der Phonerkennung, aber man erzielt auch keine Verbesserung. Sinnvoller wäre der Ansatz, die Ereignislabel unabhängig von den Phonlabeln zu gewinnen und auf die dynamische Folge erkannter Ereignisse (das Ereignisgitter) einen Phonerkenner aufzusetzen, der auch den zeitlichen Verlauf der Ereignisse berücksichtigt. Solch ein Vorgehen ist aber erst möglich, wenn es echt ereignisgelabelte Sprachdaten in größerem Umfang gibt. Beim Bayerischen Archiv für Sprachsignale (BAS) ist ein Korpus AD angekündigt,

Ereignis	falscher Alarm			nicht detektiert			Korrektheit der Entscheidung		
	1 ZS	5 ZS	Hüb	1 ZS	5 ZS	Hüb	1 ZS	5 ZS	Hüb
ap	6.2	8.2	6.7	61.5	51.1	54.5	81.2	82.0	85.1
fr	2.7	4.0	4.7	32.0	26.9	34.6	92.8	92.5	90.7
gh	2.9	3.6	3.6	28.8	22.9	22.6	93.8	93.9	94.3
gl	0.0	0.0	0.0	100	100	100	99.3	99.3	99.5
la	0.0	0.0	0.3	100	100	94.8	98.9	98.9	98.3
lb	0.8	2.1	0.4	88.3	85.1	89.2	87.4	86.7	91.1
na	4.7	4.9	1.6	32.3	29.4	62.4	91.6	91.8	92.0
op	0.9	2.5	0.2	95.9	74.3	99.6	81.3	83.6	90.8
pa	11.5	8.2	7.0	15.2	12.5	13.3	87.7	90.9	90.6
pl	1.1	1.3	1.4	53.8	51.1	42.8	97.8	97.7	97.5
po	0.0	0.0	0.3	100	94.3	37.0	99.2	99.2	99.5
sh	5.9	6.4	7.5	11.5	10.4	14.1	91.1	91.5	89.6
tv	1.0	1.1	1.1	82.5	80.4	80.9	94.4	94.4	97.1
uv	0.0	0.0	0.0	100	100	100	99.5	99.5	99.1
va	4.0	4.4	2.7	39.2	33.8	35.1	92.3	93.3	95.1
vd	5.3	6.5	4.0	27.9	22.9	22.3	90.8	90.7	93.6
ve	0.5	1.1	0.8	94.1	92.8	77.9	92.0	91.6	95.1
vg	3.1	2.6	3.8	82.7	84.7	55.6	87.1	87.3	92.4
vh	2.1	3.4	2.6	62.0	51.5	32.5	91.4	91.4	95.4
vm	0.4	1.9	5.0	95.6	82.3	59.8	90.6	90.5	90.3
vo	5.7	5.6	9.6	20.9	10.1	14.5	89.6	90.2	89.2
vr	1.9	2.1	2.1	60.0	60.0	50.5	92.0	91.9	94.1
vt	4.0	4.2	3.8	43.3	39.3	33.3	92.5	92.6	93.5
vu	5.5	6.1	8.9	39.8	35.4	25.9	87.5	88.0	88.2
vz	5.6	6.3	9.0	67.3	60.3	48.7	84.5	84.9	85.7

Tabelle 2: *Fehlerraten bzw. Erkennungsraten bei der Ereigniserkennung (in Prozent) bei den (sprecherunabhängigen) Netzen (7 - 12 - 25) (1 ZS) und (35 - 12 - 25) (5 ZS) im Vergleich zum (sprecherabhängigen) Erkenner (Hüb) von K.Hübener [2]*

der artikulatorische Daten enthält. Mit dessen Erscheinen könnten die erforderlichen Daten (wenn auch nicht genau in der hier verwendeten Form) zur Verfügung stehen. Ein Versuch, auf einem anderen Weg zu ereignisgelabelten Daten zu gelangen, wird im nächsten Kapitel beschrieben.

4 "Bootstrapping" zur Ereignislabelung

4.1 Idee

Inspiriert durch eine Arbeit von Elman und Zipser [4], bei der gezeigt wurde, daß sich in den Hidden-Schichten eines neuronalen Netzes einzelne Neuronen auf bestimmte Eingabemerkmale spezialisieren können, wurde versucht, ein ähnliches Verfahren zur Ereignislabelung anzuwenden. Elman trainierte ein neuronales Netz mit Kombinationen von Konsonanten und Vokalen auf die identische Abbildung. Er stellte dabei fest, daß bestimmte Neuronen der Hidden-Schicht auf das Auftreten bestimmter Vokale oder Konsonanten reagieren können. Die identische Abbildung bietet sich an, da sie keinerlei Labelung der Daten erfordert und somit die Menge an Trainingsdaten nahezu unbeschränkt ist. Die Aktivitäten der einzelnen Neuronen stellten sich bei Elman von selbst ein, er hatte keine zielgerichtete Initialisierung des Netzes vorgenommen. So war es nicht von vornherein absehbar, daß einzelne Neuronen der Hiddenschicht auf bestimmte Eigenschaften des Signals reagieren würden.

4.2 Experimente

Im Gegensatz zu Elman war bei uns beabsichtigt, gezielt Ereignisse auf die Hidden-Schicht abzubilden und auf diese Art und Weise zu prüfen, ob sich einzelne Ereignisse als "natürliche" Merkmale aus dem Sprachsignal stabilisieren. Dazu wäre es sehr aufwendig, die sich ausbildende Hidden-Schicht im Nachhinein auf Übereinstimmungen mit einzelnen Ereignissen zu untersuchen. Deshalb wurden die neuronalen Netze geeignet initialisiert, um die Grundrichtung der Hidden-Schicht vorzugeben. Zur Initialisierung der Netze wurden jeweils zwei Teilnetze trainiert: das erste Teilnetz sollte die Abbildung von (spektralen) Merkmalvektoren auf Ereignisse leisten und das zweite Teilnetz die dazu inverse Abbildung — von den Ereignissen zu den Merkmalvektoren. Hält man sich vor Augen, daß die Ereignisse im Idealfall binäre Vektoren sind, die Merkmalvektoren dagegen reell, ist nicht zu erwarten, daß die Abbildung von den Ereignissen auf die Merkmale befriedigende Ergebnisse liefert. Es ist nicht möglich, für die Gewichtsmatrix der Umkehrabbildung einfach die invertierte Gewichtsmatrix der Ereignisabbildung [2] zu verwenden, denn die Neuronen haben eine nichtlineare Übertragungsfunktion. Wie schon erwähnt, war nicht die Leistungsfähigkeit der Abbildungen der Teilnetze das Ziel des Trainings, sondern die grobe Initialisierung der Teilnetze. Nach der Initialisierung wurden die Teilnetze an den Ereignisschichten aneinandergefügt und das entstandene Netz mit weiterem Sprachmaterial auf die identische Abbildung trainiert. Nach Abschluß dieses Haupttrainings wurde das Netz an der Ereignisschicht wieder aufgetrennt und das neue Teilnetz auf seine Abbildungsqualität bezüglich der vortrainierten Ereignisse getestet. Dabei wurde untersucht, ob sich die Abbildungsqualität für die einzelnen Ereignisse verbessert oder verschlechtert hat.

[2] falls die Gewichtsmatrix überhaupt invertierbar ist

Als Merkmalsätze wurden Vektoren aus 8 Zeitscheiben mit jeweils 6 PLP-Koeffizienten und der Zahl der Nullstellen im Frame gewählt. Die Merkmalvektoren beginnen immer an einem Phonanfang und umfassen durch ihre Länge das ganze Phon. Das erlaubt es, in die Ereignisabbildung die Dynamik des Zeitsignals einfließen zu lassen, denn es ist jetzt nicht mehr notwendig, das Auftreten eines Ereignisses an einer bestimmte Stelle innerhalb des Phons zu labeln, sondern es wird jetzt nur das Auftreten innerhalb eines Phons an sich angezeigt. Bei gleichen Phonlabel sollte auch die zeitliche Struktur innerhalb des Phons ähnlich sein, sodaß die Bereiche mit ähnlicher Ereignisstruktur einander entsprechen. Auf diese Weise kann man das Problem der fehlenden Ereignislabelung etwas entschärfen.

Für die Experimente wurden die Zahl der Hidden-Neuronen variiert und folgende Netzgrößen gewählt:

- (56 - 25 - 56), aufgeteilt in (56 - 25) und (25 - 56) (ohne Hidden-Schicht in den Teilnetzen)
- (56 - 12 - 25 - 12 - 56), aufgeteilt in (56 - 12 - 25) und (25 - 12 - 56)
- (56 - 25 - 25 - 25 - 56), aufgeteilt in (56 - 25 - 25) und (25 - 25 - 56)

Nachdem die Teilnetze zu einem großen Netz zusammengefügt worden waren, wurden sie mit verschiedenen absteigenden Lernratenfolgen nachtrainiert. Die Folgen begannen bei den Lernraten 0.4, 0.1 und 0.02 und endeten jeweils bei einer Lernrate von 0.01. Je höher die Lernrate ist, desto geringer ist der Einfluß des Vortrainings auf die Abbildungseigenschaften der Ereignisschicht.

4.3 Interpretationsprobleme

Eine erste Auswertung der Ergebnisse zeigte, daß sich die mittlere Aktivität in der Ereignisschicht nach dem Haupttraining (A_T) deutlich höher war als nach der Initialisierung (A_I). Dies führt zu Problemen bei der Erkennung des Auftretens der Ereignisse, da dies bisher über eine universelle Schwelle von $s_0 = 0.5$ entschieden worden war. Durch eine höhere mittlere Aktivität in der Ereignisschicht ergibt sich jetzt eine entsprechend höhere Auftretenswahrscheinlichkeit für die Ereignisse. Für die Lösung dieses Problems boten sich zwei Wege an:

1. Die universelle Schwelle wird entsprechend der neuen mittleren Aktivität angehoben: $s_{neu} = s_0 \cdot \frac{A_T}{A_I}$.
2. Es wird für jedes einzelne Ereignis j eine neue, jetzt aber individuelle Schwelle s^j_{neu} bestimmt. Die individuellen Schwellen erfordern einen höheren Rechenaufwand, haben aber den Vorteil, daß man jedes Ereignis für sich optimieren kann.

4.4 Ergebnisse

Bedingt durch die Tatsache, daß hier pro auftretendem Phon nur ein Merkmalvektor generiert wird, steht für das Training nicht sehr viel Sprachmaterial zur Verfügung. Das spiegelt sich auch in den Erkennungsraten der Ereignisse wieder, denn selten auftretende Ereignisse werden deutlich schlechter erkannt als häufige Ereignisse. Relativ gut abgebildet werden nach dem Vortraining die Ereignisse "sh" (stimmhaft), "vo" (vokalisch) und "vu". Alle anderen Ereignisse wurden mangelhaft abgebildet. Es zeigte sich auch, daß bei den großen Teilnetzen ((56 - 25 - 25) und (25 - 25 - 56)) die Abbildungsleistungen durch die (zu) hohe Zahl von Parametern unbefriedigend war. Die Qualität der Umkehrabbildungen war nach dem Vortraining sehr schlecht, durch die hohe Komplexität der Aufgabe und die geringe Menge an Trainingsmaterial konvergierten die Netze auch sehr schlecht während des Trainings.

Wie schon erwähnt, stieg durch das Nachtrainieren die Aktivität in der Ereignisschicht stark an. Hätte man die Schwelle für das Auftreten eines Ereignisses auf ihrem alten Wert belassen, so wären sämtliche Erkennungsraten zufällig geworden. Der Grund für das Ansteigen der Aktivität liegt in der schlechten Leistung der Umkehrabbildungen, denn dadurch werden die beim Nachtrainieren auftretenden Fehler in der Gesamtabbildung auch in die Ereignisschicht zurückpropagiert, wodurch sich die Repräsentationen in allen Schichten ändern.

Um den Effekt der Aktivitätssteigerung in der Ereignisschicht zu kompensieren, wurde eine individuelle Schwelle für jedes Ereignis so bestimmt, daß die neue Auftretenshäufigkeit (für den Trainingsdatensatz) wieder der Häufigkeit, die durch die Label vorgegeben ist, entsprach. Die Auswertung der Ergebnisse wurde mit einem Testdatensatz durchgeführt, bei dem sich sowohl die Sprecher als auch die Äußerungen vom Trainingsdatensatz unterschieden (6 Sprecher und 2 Sprecherinnen $\rightarrow$ Test_3).

Sind beim Nachtrainieren die Lernraten und die Zahl der Iterationen klein, blieben die Abbildungen der Ereignisse auf der Ereignisschicht erkennbar. Mit wachsender maximaler Lernrate bzw. mit dem Ansteigen der Zahl von Iterationen wird aber die Bindung eines Neurons an ein bestimmtes einzelnes Ereignis (soweit sie je bestand) immer weiter aufgeweicht, bis sie schließlich nicht mehr erkennbar ist.

Die einzige Verbesserung bei der Ereigniserkennung, die bei diesem Verfahren beobachtbar war, trat bei dem Ereginis "gl" (glottal) auf. Sowohl der Fehler "falscher Alarm" als auch der Fehler "nicht detektiert" sank, wenn man die Zahl der Iterationen beim Nachtraining klein hielt. Mit steigender Zahl von Iterationen nahmen beide Fehler wieder zu. Bei den anderen Ereignissen war teilweise ein Absinken eines Fehlers zu beobachten, gleichzeitig nahm aber der andere Fehler entsprechend zu, so daß sich dieser Effekt als eine Verschiebung der Schwelle für die verschiedenen Datensätze interpretieren läßt und nicht zu einer besseren Erkennung führt.

4.5 "Bootstrapping" zur Ereignislabelung — Schlußbetrachtung

Es ist mit unserem Verfahren nicht gelungen, die Erkennung der Ereignisse zu verbessern. Dies liegt zum einen sicherlich an der Komplexität der Abbildung der Ereignisse auf die Merkmalvektoren (hier: 8 Zeitscheiben mit 6 PLP-Koeffizienten + Nullstellen), die für die identische Abbildung als Initialisierung notwendig ist (die relativ hohe Zahl von Zeitscheiben ist auf der anderen Seite notwendig, um innerhalb eines Phons die Dynamik der Ereignisse erfassen zu können). Es ist aber auch möglich, daß die zuvor definierten Ereignisse keine natürlichen Eigenschaften des Sprachmaterials sind bzw. nicht mit den hier verwendeten Merkmalen charakterisiert werden können. Dann ist auch zu verstehen, daß sich die Ereignisse nicht in der vortrainierten Schicht stabilisieren.

Unter Umständen ist auch der eingeschlagene Weg eine Sackgasse, denn in neuronalen Netzen wird die Information verteilt repräsentiert, wobei sich dann die Neuronen der Hidden-Schichten nicht mehr unbedingt auf ein (von außen vorgegebenes) Merkmal (z.B. Ereignis) spezialisieren.

In weiteren Versuchen sollte noch untersucht werden, ob sich mit den Ereignisschichten, die sich nach dem Haupttraining ergeben haben, bessere Ergebnisse bei der Phonerkennung ergeben bzw. ob eine Clusteranalyse eine bessere Interpretation der Ergebnisse erlaubt.

5 Zusammenfassung

Motiviert durch die Arbeiten K.Hübeners an artikulatorischen Ereignissen wurde an der Abteilung für Neuroinformatik der Universität Ulm versucht, die artikulatorischen Ereignisse zur Verbesserung der sprecherunabhängigen Phonerkennung heranzuziehen. Zentrale Idee für die Verwendung von Ereignissen als Merkmale ist die Tatsache, daß Ereignisse andere zeitliche Grenzen haben als Phone oder noch höhere Einheiten. Wie bekannt, sind nicht die stationären Teile eines Sprachsignals für das Verständnis von größter Bedeutung, sondern vielmehr die instationären Bereiche der Übergänge zwischen verschiedenen (stationären) Bereichen des Sprachsignals. Mit der Auflösung der Phongrenzen durch die Ereignisse und der möglichen differenziertere Beschreibung des Sprachsignals (gegenüber den Phonen) sollten sich durch die Verwendung von Ereignissen Vorteile ergeben.

Das Problem bei der Verwendung von Ereignissen ist jedoch, daß so gut wie kein echt ereignisgelabeltes Sprachmaterial zur Verfügung steht. Die Experimente wurden deshalb mit pseudo-ereignisgelabelten Daten durchgeführt, bei denen die Ereignislabelung mit Hilfe einer von K.Hübener zur Verfügung gestellten Tabelle aus der Phonlabelung gewonnen worden ist. Verloren geht dabei aber der theoretische Hauptvorteil der Ereignisse, die zeitliche Entwicklung innerhalb eines Phons zu modellieren, da die Ereignislabel jeweils für die Dauer eines Phons als konstant angesehen werden. Ungeachtet der dadurch erzwungenen Einschränkungen bei der Interpretation der Experimente können die Ereignisse dennoch als weiterer Merkmalsatz angesehen werden. Unter diesem Gesichtspunkt wurden zwei Serien von Experimenten durchgeführt:

In einer ersten Phase wurden die Ereignisse als Merkmalsatz interpretiert und versucht, durch eine zweistufige Abbildung (1.Stufe: PLP-Koeffizienten $\rightarrow$ Ereignisse, 2. Stufe: Ereignisse $\rightarrow$ Phone) die Phonerkennung zu verbessern. Im Vergleich zu einer herkömmlichen Phonerkennung mittels neuronaler Netze (feed forward) gleicher Komplexität ergaben sich durch die Verwendung der Ereignisse weder Vor- noch Nachteile bzgl. der Erkennungsraten.

In der zweiten Phase wurde versucht, mittels eines an Elman angelehnten Bootstrapping-Verfahrens mit Hilfe der identischen Abbildung eine verbesserte Ereignislabelung zu erzielen. Dabei sollten sich Ereignisse, die "natürliche" Merkmale des Sprachsignals sind, nach einer Initialisierung selbständig in der Hidden-Schicht eines neuronalen Netzes stabilisieren. Bei den Experimenten zeigte sich jedoch, daß sich kein Ereignis in Form einer Aktivität eines einzelnen Neurons in der Hidden-Schicht stabilisieren konnte. Problematisch erwies sich, daß das zur Initialisierung notwendige Teilnetz für die Abbildung von Ereignissen auf PLP-Merkmalvektoren nur sehr schlecht trainierbar ist.

Unsere Ergebnisse besagen nicht, daß das Konzept der Ereignisse nicht weiter verfolgt werden sollte. Vielmehr belegen sie die Schwierigkeiten, die die Umsetzung des Konzepts ist eine Anwendung mit sich bringen. Schon für den Menschen ist es nicht einfach, gesprochene Sprache reproduzierbar und eindeutig auf Ereignisebene zu kennzeichnen. Hinzu kommt, daß die Ereignisse nicht als akustische Merkmale definiert worden sind sondern den Artikulationsvorgang beschreiben und somit nicht zwangsläufig immer eine lokale akustische Entsprechung haben müssen. Manche Laute lassen sich auf verschiedene Arten bilden, sodaß die Zuordnung der Ereignisse zum Sprachsignal nicht eindeutig ist. Hält man sich dies vor Augen, kann man von einer Maschine, die auf den menschlichen Erfahrungen aufbaut, keine Wunder erwarten. Neben solch grundsätzlichen Problemen ist es zur Zeit auch noch unklar, welche Methoden und Sprachparametrisierungen der Aufgabe optimal angepaßt sind.

Bei unseren Experimenten haben sich die artikulatorischen Ereignisse als schwieriges Terrain erwiesen, sowohl was ihre Erkennung als auch ihre Anwendung betrifft. Ein Hauptgrund dafür ist sicherlich das Fehlen großer Sprachdatensätze, die auf Ereignisebene gelabelt sind und deren Ereignislabelung nicht starr an die Phonlabelung gebunden ist. Sollte ein solcher Datensatz zur Verfügung stehen, würde es sich lohnen zu untersuchen, ob die darin enthaltene zusätzliche Dynamik sich für die Spracherkennung sinnvoll nutzen läßt. Mit dem Erscheinen des vom Bayerischen Archiv für Sprachsignale (BAS) angekündigten Korpus AD mit artikulatorischen Daten könnte dies bald der Fall sein.

Literatur

[1] K.Hübener: *Detektion akustisch–phonetischer Ereignisse*; Verbmobil-Memo 5/93

[2] K.Hübener, J.Carson-Berndsen: *Phoneme Recognition Using Acoustic Events*; Verbmobil-Report 15, 1994

[3] H.Hermansky: *Perceptual linear predictive (PLP) analysis*; J. Acoust. Soc. Am., **87**(1990)1738-1752

[4] J.Elman, D.Zipser: *Learning the hidden structure of speech*; J. Acous. Soc. Am. **83**(1988)1615-1626

[5] A. Zell, G. Mamier, M. Vogt, N. Mache, R. Huebner, K.-U. Herrmann, T. Soyez, M. Schmalzl, T. Sommer, A. Hatzigeorgiou, S. Doering, D. Posselt, T. Schreiner: *SNNS, Stuttgart Neural Network Simulator, User manual, Version 3.2*; University of Stuttgart, Comp. Science Dept., Report No. 6/94