Enrico Biermann (enrico@cs.tu-berlin.de)
Timo Glaser (timog@cs.tu-berlin.de)
Marco Kunze (makunze@cs.tu-berlin.de)
Sebastian Nowozin (nowozin@cs.tu-berlin.de)

WS 2002/03
18. 2. 2003

YAVA - Benutzerhandbuch

1 Einleitung

Das Dokument beschreibt die Ergebnisse des Projektes "Neuronales Netz" des C++ Programmier Praktikums vom Wintersemester 2002/2003 an der TU-Berlin [2]. Dargestellt wird der theoretische Hintergrund der Anforderungen an das Programm, die Leistungsfähigkeit des Programms und die Bedienung durch den Benutzer.

2 YAVA

YAVA - Yet Another Vowel Analyzer - ist die Realisierung einer einfachen Spracherkennung. Dabei werden dem Programm Spracheingaben geliefert, die zu klassifizieren sind. Die Eingaben bestehen ausschließlich aus Vokalen, die in fünf Klassen einzuteilen sind. Intern bedient sich das Programm mehrerer neuronaler Netze, die erst trainiert und anschließend zur Klassifikation genutzt werden. Zur leichteren Bedienung verfügt das Programm über eine GUI.

3 Theoretische Grundlagen

3.1 Audio

Die menschliche Sprache erzeugt Schall, der sich etwa im Bereich von 50 bis 5000 Hertz bewegt. Dabei tragen die Vokale am meisten zum Verständnis des Gesprochenen bei, da sie am energiereichsten und harmonischten sind (stimmhaft). Aufgrund dieser sehr starken Ausprägung fällt es dem Hörer auch leichter Vokale voneinander zu unterscheiden als zum Beispiel Konsonanten wie d und t. Dabei sind vor allem die Formanten wichtig, also die Bereiche, in denen die Vokale jeweils am energiereichsten sind. So ist zum Beispiel ein u im Bereich von 1000 und 5000 Hertz sehr energiereich, wogegen ein a etwa bei 900 und 1300 Hertz Amplitudenmaxima aufweist. Beim Menschen ist für die Erkennung von Sprachlauten der Hörnerv zuständig, welcher die je nach Frequenz und Amplitude unterschiedlichen mechanischen Schwingungen der Trommelfellmembran in neuronale Entladungen umwandelt, die vom Gehirn verarbeitet werden. In unserem Projekt übernehmen diese Aufgabe neuronale Netze. Als Eingabe für die Netze empfiehlt es sich aber nicht die reinen Audiodaten zu verwenden, sondern die Energie für bestimmte Frequenzbereiche. Anhand dieser Daten lässt sich feststellen, wo sich die Formanten des jeweiligen Vokals befinden und daraus wiederum lässt sich der Vokal selbst ableiten. Näheres dazu findet man in [3] und [4].

3.2 Neuronales Netz

Eine weiter ausholende Beschreibung der in YAVA implementierten Neuronalen Netze und ihrer Lernverfahren ist in [1] zu finden.

3.3 Vokalerkennung

Die Vokalklassifizierung läuft mit Hilfe drei neuronaler Netze ab, deren Ausgaben kombiniert werden. Dabei ist jedes dieser Netzwerke ein Experte bei der Unterscheidung einzelner disjunkter Vokalgruppen. Diese sind:

Netzwerk	Gruppe 1	Gruppe 2	Gruppe 3
1	{ A }	{ O, U }	{ E, I }
2	{ O }	{ U }	{ A, E, I }
3	{ E }	{ I }	{ A, O, U }

Jedes der drei neuronalen Netzwerke besitzt drei Neuronen in der Ausgabeschicht. Alle diese Neuronen sind mit der logistischen Aktivierungsfunktion belegt, die einen Wertebereich von 0.0 bis 1.0 hat. Die Wahrscheinlichkeiten der einzelnen Vokalen wird über Produktbildung der verschiedenen Netzausgänge gebildet:

Wahrscheinlichkeit w_{vokal}	Produkt
w_a	$y_{n1,g1} \cdot y_{n2,g3} \cdot y_{n3,g3}$
w_e	$y_{n1,g3} \cdot y_{n2,g3} \cdot y_{n3,g1}$
w_i	$y_{n1,g3} \cdot y_{n2,g3} \cdot y_{n3,g2}$
w_o	$y_{n1,g2} \cdot y_{n2,g1} \cdot y_{n3,g3}$
w_u	$y_{n1,g2} \cdot y_{n2,g2} \cdot y_{n3,g3}$

Wobei die einzelnen Wahrscheinlichkeiten w_v für die Wahrscheinlichkeit steht, dass der Eingabesample einen Vokal v darstellt. Wie leicht ersichtlich, ist die maximale Wahrscheinlichkeit 1.0, nämlich genau dann, wenn sich alle drei Experten “einig sind”.

4 Technisches

Die Fähigkeiten von YAVA lassen sich ungefähr in der Schnittmenge von den Fähigkeiten von Toastern und Waschmaschinen wiederfinden. Es kann von einem Sprecher auf die Erkennung der von ihm gesprochenen Vokale trainiert werden und später zur Erkennung verwendet werden. Da die genannte Schnittmenge recht klein ist, kann man sich vorstellen, dass es viel mehr nicht von YAVA beherrschte Fähigkeiten gibt. YAVA ist nicht sprecherunabhängig, und braucht außerdem eine recht hohe Zahl an Samplen (ca. 30 pro Vokal) um trainiert werden zu können. Da keine besonderen Algorithmen zur Optimierung der Samples implementiert wurden, sollten die Samples außerdem eine recht gute Qualität haben und eine ausreichende Länge. Mit einem gut eingestellten Mikrofon, und wenn YAVA zur Aufnahme verwendet wird sollten die Samples gut genug sein.

Hardwaretechnisch ist YAVA der Traum jedes Low-Budget-Unternehmens. Man braucht nicht mehr als ein Mikrofon (damit natürlich auch eine Soundkarte) und ein Linux oder Solaris. Rechenleistung wird nur zum Trainieren gebraucht, die danach an den Prozessor gestellten Aufgaben sind gering.

Auch die Installation von YAVA ist unproblematisch. Die Quelltexte befinden sich in `./src`. Eventuell muss das `Makefile` angepasst werden, da die Compilerversionen und die Pfade zu den QT-Libraries zwischen den Systemen variieren können (Für nähere Informationen wenden Sie sich einfach an Ihren Systemadministrator, die freuen sich wenn jemand mit ihnen redet). `make` kompiliert das Projekt, die ausführbare Datei ist `Yava`, die sich in `./src/gui` befindet. Außerdem kann der Splashscreen variabel gewählt werden, indem eine beliebige png-Datei nach `./src/gui/yava.png` kopiert wird. Drei fertig designte Bilder (`./src/gui/yava_[1-3].png`) sind bereits mitgeliefert.

5 Benutzeroberfläche

Die GUI von Yava wurde über ein Tab-Widget realisiert, auf dessen Seiten alle Module des Programms konfiguriert und gesteuert werden können. Es handelt sich dabei um jeweils eine Seite für die Audiovorverarbeitung, den Setmanager, die Konfiguration und Erstellung des Neuronalen Netzes, die Einstellung der Lernparameter und Starten des Lernvorgangs, und zu guter Letzt eine Seite, auf der die Ausgaben des Neuronalen Netzes und das Neuronale Netz selbst dargestellt werden. Die Tab-Reihenfolge stimmt quasi mit den Bedienungsschritten überein, um eine intuitive Benutzerführung zu gewährleisten.

5.1 Die Audiovorverarbeitung

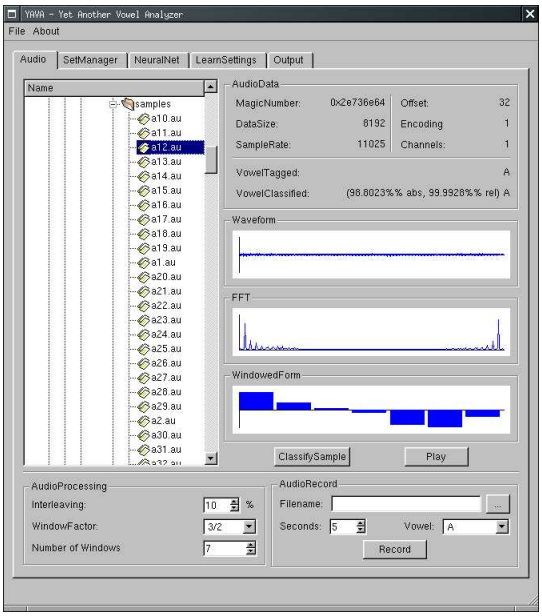

Abbildung 1: Audiovorverarbeitung

Die Audio-Seite (Abbildung 1) ermöglicht das Betrachten einzelner Dateien, Einstellungen zur Audiovorverarbeitung und das Aufnehmen der Dateien.

Mit dem Dateibrowser kann der Benutzer Audio-Dateien auswählen um sich Informationen über diese anzeigen zu lassen. Diese bestehen aus Informationen wie Aufnahmefrequenz, Länge der Datei, etc. Neben den reinen Audioinformationen wird aber auch der in der Datei enthaltene Vokal, sowie, wenn denn bereits ein Neuronales Netz initialisiert wurde, der erkannte Vokal dargestellt. Zusätzlich wird die Datei grafisch als gewohnte Waveform dargestellt, darunter das Ergebnis der FFT-Analyse und schließlich die Fensterung, die als Eingabewert des Neuronalen Netzes dient.

Unten Links kann der Benutzer Einstellungen für die Vorverarbeitung der Audiodaten vornehmen. “Interleaving” gibt dabei an, wie weit sich die Fenster überschneiden sollen, “Factor” gibt den Größenfaktor der Fenster an. Außerdem kann die Anzahl der Fenster eingestellt werden. Dieser Wert hat sofortige Auswirkung auf das Neuronale Netz, da die Anzahl der Neuronen in der Eingabeschicht immer der Anzahl der Fenster entsprechen muss.

Desweiteren bietet die Audio-Seite die Möglichkeit zur Aufnahme neuer Audiodateien. Anzugeben ist dabei der Dateiname, die Länge der Datei und der aufzunehmende Vokal.

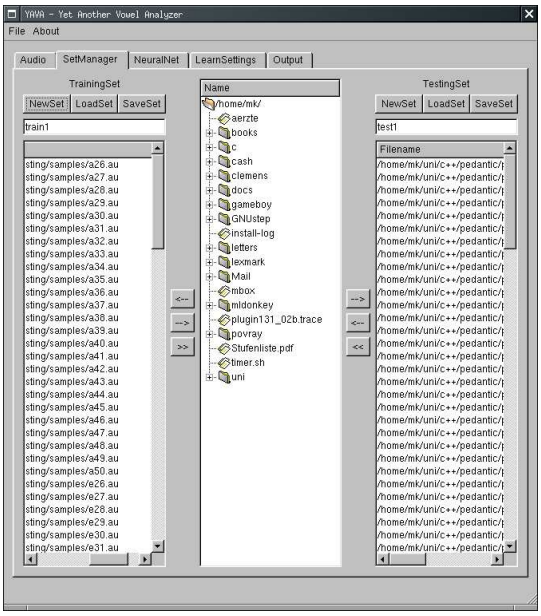

Abbildung 2: Der Setmanager

5.2 Der Setmanager

Der Setmanager (Abbildung 2) liefert alle Möglichkeiten, die man zum Erstellen und Bearbeiten von Sets braucht. Aus dem in der Mitte platzierten Dateibrowser können Dateien in das Trainingset (links) oder Testset (rechts) mit den Buttons zwischen den Listen aufgenommen werden. Der unterste der Buttons ist dafür da, die Dateien von einem Set in das gegenüberliegende zu verschieben. Selbstverständlich können die fertig zusammengestellten Sets auch abgespeichert und geladen werden.

Für gute Ergebnisse bei der Vokalerkennung sollten die Sets mindestens 3 Vokale jeder Art enthalten, die Sets sollten außerdem disjunkt sein, da sonst die Fehlerkurven nicht korrekt sind (Das Netz kann eine Audiodatei mit der es trainiert wurde natürlich sehr viel besser klassifizieren als andere).

5.3 Das Neuronale Netz

Nachdem die Sets eingerichtet wurden, muss man das Neuronale Netz einrichten und konfigurieren (Abbildung 3). YAVA benutzt zur Klassifizierung der Vokale intern drei Neuronale Netze. Ein Tab-Widget ermöglicht die Konfiguration der Netze unabhängig voneinander. Der Benutzer hat dabei die Möglichkeit, die Anzahl der Schichten mit den dazugehörigen Neuronenzahlen einzustellen, ebenso die Aktivierungsfunktion für der Eingabeschicht und für die versteckten Schichten. Wir die Anzahl der Neuronen in der Eingabeschicht gespeichert, so wird der Wert auch gleichzeitig für die Fensterung überommen, da die beiden Werte gleich sein müssen. Die Initialisierung der Gewichte und Schwellwerte wird über einen Zahlenbereich angegeben, aus dem dann die Anfangswerte per Zufallsfunktion ermittelt werden.

Um ein Netz letztendlich zu benutzen, muss es noch generiert, was über den Button “GenerateSystem” geschieht. Es besteht auch die Möglichkeit Netze zu speichern und zu laden. Wird ein Netz geladen ist es automatisch generiert. Beim Abspeichern werden auch die aktuellen Gewichte und Schwellwerte berücksichtigt, es können also auch angelernte Netze korrekt abgespeichert und geladen werden. Ebenso sind die Einstellungen der Audiovorverarbeitung in der Datei mit abgespeichert.

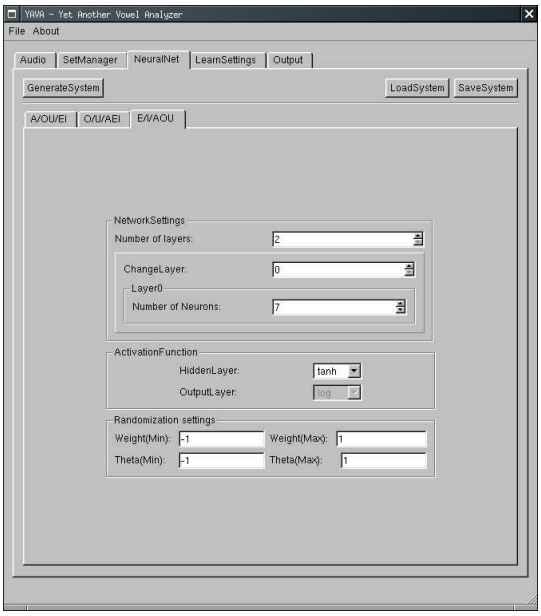

Abbildung 3: Das Neuronale Netz

5.4 Die Lerneinstellungen

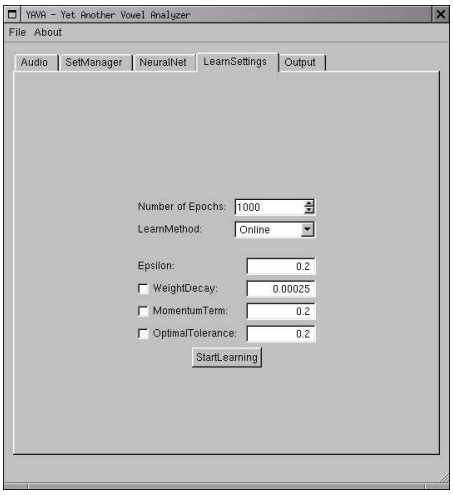

Abbildung 4: Die Lerneinstellungen

Die Lerneinstellungen (Abbildung 4) sind zur Feinabstimmung des Lernvorgangs.

Die Anzahl der Epochen bestimmt, wie häufig das Netz mit den beiden Sets trainiert werden soll, Online- und Batchlearning sind zwei verschieden Lernmethoden, deren Bedeutung der Literatur [2] entnommen werden kann. Der interessierte Benutzer informiere sich auch über die vier anderen Werte und deren Bedeutung aus der einschlägigen Literatur, die Standardwerte sind bereits annehmbar.

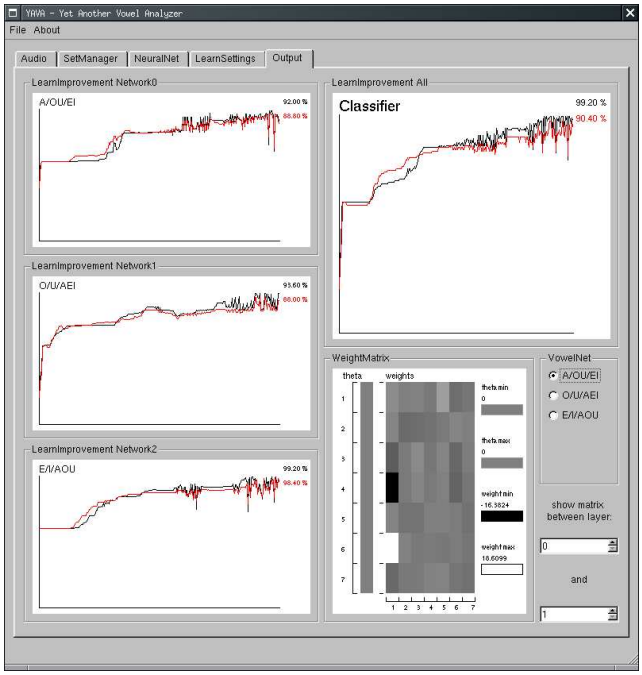

Abbildung 5: Die Ausgabe

5.5 Die Ausgabe

Zu guter Letzt kann man sich einen Überblick (Abbildung 5) über die Netze und den Lernerfolg verschaffen. Dazu werden die Lernkurven der drei einzelnen Netze dargestellt, sowie deren kumulativer Lernerfolg. Die Netze selbst kann man sich über eine Gewichtsmatrix anzeigen lassen. Der Benutzer kann hierzu das anzuzeigende Netz und die beiden Schichten, deren zwischenliegende Gewichte angezeigt werden sollen, auswählen.

Literatur

[1] pedantic project team, “Grundlagen neuronaler Netzwerke”

[2] C++-Programmier-Praktikum, “Projektaufgabe Pj-NN”
`http://ni.cs.tu-berlin.de/lehre/C-Praktikum/data/Projekte/projekt_nn.pdf`

[3] Vokalklassifikation mit Serial Competition Vektorquantisierung
`http://www-lehre.informatik.uni-osnabrueck.de/ nn/Praktikum/projekte/beatefrank/`

[4] Diplomarbeit Stefan Bleeck
`http://www.tonhoehe.de/diplom-node1.html`