- Studienarbeit - Automatische Bestimmung von Visemen für das maschinelle Lippenlesen

Rainer Stiefelhagen

Betreuer: Dr. Paul Duchnowski

Institut für Logik, Komplexität und Deduktionssysteme
Lehrstuhl Professor A. Waibel
Fakultät für Informatik
Universität Karlsruhe (TH)
D-76128 Karlsruhe

12. April 1995

Inhaltsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1 Lippenlesen

Derzeitige Spracherkennungssysteme arbeiten meist nur dann robust, wenn sie in einer geräuschfreien Umgebung, wie etwa einem Labor, benutzt werden. Treten Hintergrundgeräusche, Zwischenrufe etc. auf, sinkt die Erkennungsrate dieser Systeme rapide ab.

Eine Möglichkeit, diesen Effekt zu mildern bietet die zusätzliche Nutzung visueller Information. So können Bilder der Lippenbewegung des Sprechers nützliche zusätzliche Information zur Spracherkennung bieten. Verschiedene Arbeiten auf diesem Gebiet [8, 15, 17, 18, 22] haben gezeigt, daß kombinierte audio-visuelle Spracherkennung möglich ist und Fortschritte in der Erkennungsleistung erzielt werden können.

Auch an der Universität Karlsruhe wird an einem Spracherkennungssystem gearbeitet, das zusätzliches Lippenlesen zur Verbesserung der Erkennungsleistung benutzt. Dazu werden mit einer herkömmlichen Videokamera Aufnahmen des Sprechers gemacht, und diese — nach einigen Vorverarbeitungsschritten — als visuelle Eingabe für das System verwendet [4, 5, 13]. Dieses automatische Spracherkennungsystem soll im Moment kontinuierlich gesprochene deutsche Buchstaben erkennen und ist derzeitig noch sprecherabhängig. Das Erkennungssystem besteht im wesentlichen aus zwei TDNN's: eines zur Erkennung von Phonemen und eines zur Erkennung von Visemen. Die Ausgaben der einzelnen TDNN's werden dann in einer gemeinsamen Schicht kombiniert und als Eingabe für einen DTW-Algorithmus verwendet.

Genaue Beschreibungen des akustischen Erkenners finden sich in [9, 10]. Beschreibungen des kombinierten visuellen und akustischen Erkenners finden sich in [1, 2]

1.2 Viseme

Aus Forschungsarbeiten mit Hörgeschgeschädigten ist bekannt, daß sich verschiedene englische Konsonanten und Vokale nicht oder fast nicht visuell unterscheiden lassen. So lassen sich zum Beispiel die Phoneme /p/,/b/,/m/ oder /f/ und /v/ visuell nicht unterscheiden [20, 14, 21, 11, 6] .

Solche Phoneme, die das gleiche visuelle Aussehen haben und daher stark miteinander verwechselt werden, können zu einer Gruppe, einem sogenannten Visem, zusammengefaßt werden. Diese zusammengefaßten Gruppen (Viseme) lassen sich gut voneinander unter-

scheiden. Viseme bezeichnen also visuell unterscheidbare Spracheinheiten, anolog zu den Phonemen als akustisch Unterscheidbare Spracheinheiten. Eine ausführliche Erklärung findet sich zum Beispiel in [11].

Ein Kriterium für das zusammenfassen von einzelnen Phonemen zu Visemen ist eine korrekte Zuordnung dieser Phoneme zu der Visemgruppe von mindestens 70-75% (sogenannte "ingroup classification rate") [11]. So zählt beispielsweise eine visuelle Erkennung des Phonems /p/ bei tatsächlich gesprochenem /b/ als korrekte Erkennung der Visemklasse (/p/,/b/,/m/). Diese Gruppen können sich je nach Art des Sprechers, des phonetischen Kontextes, und dem Vermögen des Lippenlesers etwas unterscheiden [11, 20, 14]. Außerdem kann es Phoneme geben, die keinem Visem zugeordnet werden können [14]. Walden,Prosek,Montgomery,Scherr und Jones [20] geben zum Beispiel folgende neun Viseme für Konsonanten an: (/p/,/b/,/m/), (/r/), (/f/,/v/), (/w/), (/th/,/dh/), (/d/,/g/,/j/,/k/,/n/,/t/), (/l/), (/s/,/z/) , (/sh/,/zh/).

Jeffers und Barley [12] geben folgende acht Viseme für Vokale an: (/u/,/U/,/oU/,/δ/), (/i/,/I/,/eI/,/Λ/), (/$\supset$/), (/aU/), (/ε/,/æ/,/a/), (/$\supset$I/), (/aI/).

Da nun bekannt ist, daß sich bestimmte Phoneme auch von geübten Lippenlesern (zum Beispiel hörbehinderten Menschen) nicht unterscheiden lassen, kann man davon ausgehen, daß diese Unterscheidung auch für ein automatisches Spracherkennungssystem rein visuell nicht möglich ist. Folglich sollte man auch das dem Erkenner zugrundeliegende Neuronale Netz nicht zwingen, ununterscheidbare Phoneme zu klassifizieren. Würde man das Netz dazu zwingen, so könnte es versuchen, irrelevante Merkmale der visuellen Eingabedaten zu lernen und damit eine schlechtere Erkennungsrate auf neuen Testdaten zu erzielen.

Bisher wurde das Netz darauf trainiert, einen Satz von 42 Visemen zu klassifizieren. Diese Viseme wurden von Hand ausgewählt und optimiert [3]. Die Auswahl der Visemgruppen orientierte sich an der bekannten Literatur über englische Viseme. Anhang A.2 zeigt die für diesen Task verwendeten Phoneme, Anhang A.3 zeigt die von Hand bestimmte Einteilung in 42 Viseme.

1.3 Ziel der Studienarbeit

Ziel dieser Studienarbeit ist es, solche visuell stark verwechselbaren Phoneme automatisch zu bestimmen und zu geeigneten Klassen zusammenzufassen. Als Maß für eine "gute" Klasseninteilung dient hierbei die höchste kombinierte Erkennungsrate (Word Accuracy). Im folgenden werde ich diese zusammengefaßten Klassen auch als Viseme bezeichnen, obwohl die zusammengefaßten Phoneme nicht immer das Kriterium der mindestens 70%-igen korrekten Zuordnung zur Visemklasse erfüllen. Dies liegt aber in erster Linie daran, daß das automatische Lippenlesen nicht die Erkennungsleistung des Menschen erreicht.

2. Verwendete Datenmenge

Als Bilddatenmenge standen insgesamt 197 Bildsequenzen eines männlichen Sprechers, der sowohl verschiedene Namen, als auch sinnlose Buchstabenfolgen, buchstabiert, zur Verfügung [1]. Zum Training und Testen des Erkenners wurden sowohl Graustufenbilder der Größe 24 x 16 Pixel verwendet, die vorher grauwertequalisiert und automatisch zentriert wurden [13], als auch "LDA-Bilder" dieser Grauwertbilder [13] .

Diese "LDA-Bilder" wurden mittels Linear Discriminant Analysis (LDA) [7] aus den Grauwertbildern gewonnen. LDA ist ein Verfahren zur Dimensionsreduktion der Merkmalsvektoren, bei dem die Trennbarkeit der einzelnen Klassen (z. B. Phoneme) optimiert wird. Hier wurden Bilder mit 32 LDA-Koeffizienten benutzt.

Als Grundlage für die akustische Erkennung standen die entsprechenden FFT-Daten der 197 Sequenzen zur Verfügung (16 MelScale Koeffizienten).

Abbildung 2.1 zeigt einige (nicht vorverarbeitete) Aufnahmen einer Bildsequenz.

2.1 Trainings- und Testdatenmenge

Als Trainingsdatenmenge für die visuelle Erkennung wurden 137 Bildsequenzen benutzt, als Cross-evaluation-Menge und als Testmenge wurden jeweils 30 Bildsequenzen verwendet. Die akustischen Daten wurden entsprechend aufgeteilt.

Kurz vor Beendigung der Studienarbeit stellte sich heraus, daß eine der dreißig FFT-Dateien der Testdatenmenge fehlerhaft war. Dies hatte zur Folge, daß diese Sequenz akustisch nie erkannt wurde, und dadurch insgesamt die Erkennungsrate etwas schlechter wurde. Alle Ergebnisse, die in dieser Studienarbeit genannt werden beziehen sich auf diese Datenmenge mit einer fehlerhaften FFT-Datei.

Da sich alle in dieser Studienarbeit genannten Ergebnisse auf diese "fehlerhafte" Datenmenge beziehen, sind diese trotzdem repräsentativ und können Unterschiede der verschiedenen Verfahren aufzeigen.

[1] Bildsequenzen aus den Datenbanken mum21, mum22 und mum23

2.2 Bilddaten zur Bestimmung der neuen Visemklassen

Zur Zusammenfassung visuell ähnlicher Phoneme wurden hier die Graustufenbildsequenzen verwendet.

Die vorhandenen 197 Bildsequenzen eines Sprechers wurden anhand akustischer labels in Bildsequenzen einzelner Phoneme segmentiert. Es standen damit zu jedem der 63 Phoneme eine Anzahl von entsprechenden Grauwertbildsequenzen zu Verfügung, die dann für die verschiedenen hierarchischen Clusteringverfahren (siehe Kapitel 3 und 4) verwendet wurden.

Das Erkennungssystem wurde dann, nach Vorliegen dieser neuen Viseme, sowohl mit Grauwertbildern als auch mit LDA-Bildern trainiert und getestet.

2.3 Zugrundeliegende Phoneme

Anhang A.2 zeigt die für das Erkennenen von Buchstaben verwendeten Phoneme. Anhang A.1 zeigt die Beschreibung der einzelnen Buchstaben durch diese Phoneme.

Die Bilddaten für "silence", also für die Phoneme /si1/ und /si2/, sowie die Bilddaten für "glottal stop", also das Phonem /gs/, wurden nicht mit ins Clustering einbezogen. Erstere nicht, wegen ihrer hohen apriori-Wahrscheinlichkeit, letzter nicht. da angenommen werden kann, daß ein "glottal stop" nicht anhand der Lippenbewegungen erkannt werden kann. Es wurden also bei den einzelnen Ansätzen nur die übrigen sechzig Phoneme geclustert.

2.4 Notation

Ein Graustufenbild B_j der Größs 24 x 16 pixel wird im Folgenden als ein 384-dimensionaler Merkmalsvektor $\vec{b}_j$ aufgefaßt. Die Intensitäten der entsprechenden Pixel von $\vec{b}_j$ werden dann mit $\vec{b}_{j,i}$ bezeichnet, mit $0 < i \leq 384$. Eine Bildsequenz S entspricht einer Folge von Einzelbildern: $S = \vec{b}_1, \vec{b}_2, \ldots, \vec{b}_n$.

Abbildung 2.1: Beispiele aus der Datenbank mum21

3. Clustern mit Mittelwertbildern

3.1 Idee

Bei diesem ersten einfachen Ansatz, verschiedene Bildsequenzen hierarchisch zu clustern, wurde davon ausgegangen, daß sich Ähnlichkeiten und Unterschiede von Bildsequenzen verschiedener Phoneme auch dann noch erkennen lassen, wenn die Bildsequenzen der einzelnen Phoneme jeweils durch ein sogenanntes Mittelwertsbild repräsentiert werden. Dieses Mittelwertbild wird durch Mittelung über alle Bilder einer Klasse berechnet. Information über die zeitliche Abfolge der Bilder einer Bildsequenz werden bei diesem Verfahren also nicht mitberücksichtigt.

3.2 Mittelwertbilder

Hier wurde also zuerst von von jeder Bildsequenz $B = \vec{b}_1, \ldots, \vec{b}_n$ ein Mittelwertsbild $\vec{s}$ berechnet:

$$\vec{s}_i = \frac{\sum_{j=1}^{n} \vec{b}_{j,i}}{n}, 0 \leq i < 384$$

Danach wurde über alle m Mittelwertsbilder $\vec{s}$ einer Klasse $\mathcal{C}_p$ ein neues Mittelwertsbild $\vec{c}_j$ berechnet:

$$\vec{c}_{j,i} = \frac{\sum_{j=1}^{m} \vec{s}_{j,i}}{m}, 0 \leq i < 384$$

Jede Klasse $\mathcal{C}_p$ wird nun also durch ihr Mittelwertsbild $\vec{c_p}$ repräsentiert. Als Distanzmaß zwischen zwei Klassen wurde der Mittlere Quadratische Abstand (MSE) zwischen den Mittelwertbildern dieser Klassen gewählt:

$$D(\mathcal{C}_k, \mathcal{C}_l) = D(\vec{c_k}, \vec{c_l}) = \sum_{i=1}^{384} (\vec{c}_{k,i} - \vec{c}_{l,i})^2$$

3.3 Algorithmus

Die 60 Phonemklassen wurden nun folgendermaßen Zusammengefaßt: (Die Phoneme /si1/,/si2/ und /gs/ wurden nicht mitberücksichtigt (siehe 2.3)

1. jedes der 60 Phoneme entspricht einer Klasse, Anzahl der Klassen ist 60
2. berechne zu jeder Klasse $\mathcal{C}_p$ deren Mittelwertbild $\vec{c}_p$.
3. berechne Distanz zwischen allen Klassen: $D(\mathcal{C}_k, \mathcal{C}_l), k \neq l, 0 \leq k, l < \#Klassen.$
4. Vereinige die Klassen $\hat{\mathcal{C}}_k, \hat{\mathcal{C}}_l$ mit minimaler Distanz.
5. Solange die Anzahl der Klassen größer als Abbruchanzahl, gehe zu 2.
6. Ende

Es wurde dann bis auf verschiedene Anzahlen von Klassen (Visemen) geclustert. Anschließend wurde mit der jeweiligen Anzahl von Visemen der Erkenner neu trainiert, und die Erkennungsrate ermittelt, um eine optimale Anzahl von Klassen zu finden.
Die Reihenfolge der einzelnen Zusammenfassungen zeigt Anhang B

3.4 Ergebnisse

Der Erkenner wurde mit 23 Visemen, 32 Visemen und 42 Visemen trainiert. Tabelle 3.1 zeigt die rein visuellen Erkennungsraten mit den jeweiligen Visemen und zum Vergleich die visuelle Erkennungsrate mit der bereits vorhandenen, "von Hand" bestimmten Einteilung in 42 Viseme, sowie die visuelle Erkennungsrate ohne Verwendung von Visemen, also bei Klassifikation aller Phoneme.

#Viseme	Verfahren	WordAccuracy
23	MW-Bilder	10.3 %
32	MW-Bilder	18.2 %
42	MW-Bilder	14.6 %
(63)	ohne Viseme	28.5 %
42	von Hand	29.7 %

Tabelle 3.1: Visuelle Erkennungsraten mit Mittelwert-Clustering

Da die Einteilung mit Hilfe dieses Ansatzes im Vergleich zu den "handgemachten" Visemen offensichtlich sehr viel schlechtere Ergebnisse lieferte, wurden keine weiteren Ergebnisse mit diesem Ansatz untersucht.

4. Vergleich der Bildsequenzen mittels dynamischer zeitlicher Anpassung

4.1 Idee

Um die zeitliche Information der Bildsequenzen beim Clustern mitzuberücksichtigen genügt es nicht, über die Bilder einer Sequenz zu mitteln. Es müssen die ganzen Bildsequenzen miteinander verglichen werden.

4.2 Dynamic Time Warping

Um Bildsequenzen verschiedener Länge miteinander vergleichen zu können, müssen die einzelnen Bilder der Sequenzen aufeinander abgebildet werden. Ein übliches Verfahren, solche Sequenzen optimal aufeinander abzubilden, ist der Dynamic-Time-Warping-Algorithmus.

4.2.1 Distanz zweier Bildsequenzen

Die Bildsequenzen A,B können als Sequenzen von 384-dimensionalen Merkmalsvektoren aufgefaßt werden (siehe Kapitel 2.4).

Die zeitliche Differenz zweier Sequenzen kann durch ein Reihe von Punkten $c = (i, j)$ beschrieben werden:

$$F = c(1), c(2), \ldots, c(k), \ldots, c(K),$$

mit

$$c(k) = (i(k), j(k)).$$

Diese Folge realisiert sozusagen eine Abbildung der Zeitachse von Sequenz A auf Sequenz B. F wird als Abbildungsfunktion (warping function) bezeichnet.

Als Maß für die Distanz zweier Vektoren (Bilder) wurde hier wieder die Mittlere Quadratische Distanz genommen:

$$d(c) = d(i,j) = \frac{1}{2}\sum_{l=1}^{384}(a_i[l] - b_j[l])^2$$

Die gewichtete Summe der Distanzen auf der Abbildungfunktion F ist dann:

$$E(F) = \sum_{k=1}^{K} d(c(k)) \cdot w(k),$$

wobei $w(k) \geq 0$ ist. Hier waren die Gewichte $w(k)$ konstant: $w(k) = 1, 0 < k \leq K$.

Dieses Abstandsmaß $E(F)$ wird minimal, wenn die Abbildungsfunktion F so gewählt wird, daß sie die Zeitunterschiede der Sequenzen optimal anpaßt.

Diese optimale Anpaßung der Funktion F kann mit dem DTW-Algorithmus erreicht werden. Als Ergebnis liefert der Algorithmus die minimale Distanz zweier Bildsequenzen:

$$D(A,B) = Min_F \left[\frac{\sum_{k=1}^{K} d(c(k)) \cdot w(k)}{\sum_{k=1}^{K} w(k)} \right].$$

Eine genaue Beschreibung des DTW-Algorithmus' findet sich zum Beispiel in [16].

4.2.2 Distanz zweier Klassen

Um die Distanz zweier Klassen $\mathcal{C}_i, \mathcal{C}_j$ zu berechnen, werden mittels des DTW-Algorithmus' die Abstände zwischen jeder Bildfolge A_m aus Klasse $\mathcal{C}_i$ und jeder Bildfolge B_n aus Klasse $\mathcal{C}_j$ berechnet, aufaddiert und mit der Gesamtanzahl der Vergleiche normiert:

$$\bar{D}(\mathcal{C}_i, \mathcal{C}_j) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} D(A_m, B_n)}{m \cdot n},$$

wobei M und N die Anzahl der Bildfolgen in den Klassen C_i und C_j bezeichnen.

4.3 Gesamter Algorithmus

1. Anfangs 60 Klassen entsprechend den Phonemen (siehe Kapitel 2.3)
2. Berechne Distanzen zwischen allen Klassen mittels DTW-Algorithmus
3. Fasse Klassen mit minimaler Distanz zusammen
4. Falls die Anzahle der Klassen größer als Abbruchanzahl ist, gehe zu 2.
5. Ende

Es wurde hier bis zu einer Anzahl von zehn Klassen geclustert, und danach der Erkenner mit verschiedenen Anzahlen von Visemen trainiert, um die optimale Anzahl von Visemklassen zu finden.

Anhang C zeigt die Reihenfolge der einzelnen Klassenzusammenfassungen und eine Einteilung in 41 Viseme.

4.4 Anmerkung zur Laufzeit des Verfahrens

Um die anfangs dreiundsechzig Klassen bis auf zehn Klassen zusammenzufassen, lief das Verfahren eine Woche auf einer Workstation (DEC 3000/600). Dies liegt an der sehr hohen Anzahl von Vergleichen von Bildsequenzen, um die Distanz zweier Klassen zu bestimmen.

4.5 Ergebnisse

4.5.1 Rein visuelle Erkennung

Tabelle 4.1 zeigt die rein visuelle Erkennungsrate (Word-Accuracy) mit verschiedenen Visemklassen auf Graustufenbildern und "LDA-Bildern". Auch hier wurde wieder das Ergebnis mit von Hand ausgewähleten Visemen und bei direkter Klassifikation aller Phoneme hinzugefügt.

Verfahren	#Viseme	Graustufenb.	LDA
DTW-clustering	31	16.4 %	28.5 %
DTW-clustering	42	24.9 %	46.7 %
DTW-clustering	52	30.9 %	50.3 %
DTW-clustering	57	20.0 %	-
ohne Viseme	(63)	28.5 %	55.2 %
von Hand	42	29.7 %	52.7 %

Tabelle 4.1: visuelle Erkennungsraten mit DTW-Clustering

Man sieht, daß hier mit einer Einteilung der Phoneme zu 52 Visemen, das beste Ergebnis erzielt wurde. Bei der Erkennung mit Graustufenbilder konnt das "von Hand" erzielte Referenzergebnis leicht übertroffen werden, bei der Erkennung mit LDA-Bildern wurde das Referenzergebnis fast erreicht. Es läßt sich außerdem erkennen, daß das Ergebnis bei einer Erhöhung der Visemklassenanzahl auf 57 Klassen, wieder schlechter wird.

4.5.2 Kombinierte Erkennung

Tabelle 4.2 zeigt die Ergebnisse der Erkennung mit visueller und akustischer Eingabe. Zur visuellen Erkennung wurden hier die Graustufenbildsequenzen verwendet. Es wurde die Erkennung auf unverrauschten akustischen Daten und künstlich verrauschten akustischen Daten untersucht. Dabei wurden die akustischen Daten mit weißem Rauschen in zwei verschiedenen Särken, 16 db SNR und 8 db SNR, verrauscht.

Hier wurden die besten Ergebnisse mit automatisch bestimmten Visemen bei einer Einteilung in 52 (bei 8 db SNR) bzw. in 57 Viseme (unverrauscht) erzielt, wobei bei mit den

Verfahren	#Viseme	unverrauscht	16 db SNR	8 db SNR
DTW-clustering	31	93.9 %	67.3 %	44.2 %
DTW-clustering	42	95.2 %	69.1 %	49.1 %
DTW-clustering	52	95.8 %	72.1 %	52.1 %
DTW-clustering	57	96.4 %	69.7 %	49.1 %
ohne Viseme	(63)	96.4 %	70.3 %	50.3 %
von Hand	42	95.8 %	73.9 %	50.3 %
rein akustisch	-	93.9 %	64.8 %	39.4 %

Tabelle 4.2: kombinierte Erkennungsraten mit DTW-Clustering und Eingabe von Graustufenbildern

schwach verrauschten Daten (16 db SNR), das Referenzergebnis nicht ganz erreicht wurde. Allerdings wurde bei den stärker verrauschten Daten (8 db SNR) ein etwas besseres Ergebnis erzielt. Außerdem wurden sogar mit nur 31 Visemen noch recht gute Erkennungsraten erzielt. Hier zeigt sich auch daß hier ohne Verwendung von Visemen, bei direkter Klassifikation aller Phoneme, sehr gute Ergebnisse erzielt werden.

Tabelle 4.3 zeigt die entsprechenden kombinierten Ergebnisse bei Verwendung von LDA-Bildern als Eingabe.

Verfahren	#Viseme	unverrauscht	16 db SNR	8 db SNR
DTW-clustering	31	93.9 %	70.9 %	52.7 %
DTW-clustering	42	95.2 %	73.9 %	60.6 %
DTW-clustering	52	95.2 %	72.7 %	58.8 %
von Hand	42	96.4 %	77.0 %	63.0 %
rein akustisch	-	93.9 %	64.8 %	39.4 %

Tabelle 4.3: kombinierte Erkennungsraten mit DTW-Ansatz und Eingabe von LDA-Bildern

Hier wurde das Referenzergebnis nur fast erreicht, wobei mit der Einteilung in 42 Viseme etwas bessere Ergebnisse als mit 52 Visemen erzielt wurden.

5. Zusammenfassung der Phoneme anhand einer Konfusionsmatrix

5.1 Idee

Es werden hierbei nicht die Bildsequenzen nach ihrer Ähnlichkeit zusammengefaßt, sondern es werden solche Phoneme zusammengefaßt, die vom Neuronalen Netz des Erkenners, bei rein visueller Eingabe, stark miteinander verwechselt wurden.

Dabei handelt es sich um ein analoges Vorgehen wie bei der Forschung mit Hörgeschädigten: Dort werden den Lippenlesern eine Reihe von Buchstaben als Stimuli präsentiert, und dann deren Antworten, beziehungsweise Ergebnisse, aufgezeichnet. Hieraus ergibt sich eine Konfusionsmatrix der Phoneme. Anhand dieser Matrix läßt sich nun erkennen, welche Phoneme stark miteinander verwechselt wurden. Diese können dann sukzessive zu Klassen zusammengefaßt werden [20, 21, 14].

Hier wurde untersucht, welche Phoneme, das Neuronale Netz bei rein visuellem Training miteinander verwechselt. Auch hier wurde dazu automatisch eine Konfusionsmatrix des Erkenners auf Phonemebene erstellt, die zeigt, wie oft bei welchen Stimuli, welche Phoneme erkannt wurden.

5.2 Erzeugung der Konfusionsmatrix des Neuronalen Netzes

Zur Erzeugung der Konfusionsmatrix M des Neuronalen Netzes wurden bei einem Testlauf auf den Trainingsdaten während jedes Eingabe-Frames das tatsächliche Eingabephonem und das erkannte Phonem aufgezeichnet und ein Eintrag an der entsprechenden Stelle in der Matrix erhöht.

5.3 Auswertung der Konfusionsmatrix

Sei M[i][j] die Anzahl der Phonem P_j-Klassifikationen bei Eingabe des Phonems P_i und $Summe(P_x)$ die Gesamtanzahl der Eingaben von Phonem P_x. Dann läßt sich der Grad der Verwechslung zweier Phoneme V von P_i und P_j, folgendermaßen bestimmen:

$$V(P_i, P_j) = \frac{M[i][j] + M[j][i]}{Summe(P_i) + Summe(P_j)}.$$

Die Verwechslung V berechnet sich also aus der Summe der Verwechslungen von P_i mit P_j und der Verwechslungen von P_j mit P_i, normiert mit der Summe Eingaben der beiden Phoneme.

Die Verwechslung zweier Visemgruppen läßt sich folgendermaßen berechnen.

$$V(\mathcal{M}, \mathcal{N}) = \frac{\sum_{P_i \in \mathcal{M}} \sum_{P_j \in \mathcal{N}} M[i][j] + \sum_{P_j \in \mathcal{M}} \sum_{P_i \in \mathcal{N}} M[j][i]}{\sum_{P_i \in \mathcal{M}} Summe(P_i) + \sum_{P_j \in \mathcal{N}} Summe(P_j)}$$

$V(\mathcal{M}, \mathcal{N})$ berechnet sich also aus der Verwechslung aller Phoneme P_i aus $\mathcal{M}$ mit allen Phonemen P_j aus $\mathcal{N}$ und umgekehrt, normiert mit Summe aller Vorkommen der Phoneme aus $\mathcal{M}$ und $\mathcal{N}$.

Dieses Maß für den Grad der Verwechslung ist symmetrisch, das heißt, $V(\mathcal{M}, \mathcal{N}) = V(\mathcal{N}, \mathcal{M})$.

5.4 Gesamtes Vorgehen

Hier das gesamte Vorgehen im Überblick:

1. Starte Testlauf des Neuronalen Netzes auf Trainingsdaten und berechne Konfusionsmatrix
2. Auswertung der Konfusionsmatrix:
   a) Seien anfangs 60 Klassen entsprechend den Phonemen definiert:

      $$\mathcal{C}_i = \{P_i\}, 0 < i \leq 60$$

   b) berechne Grad der Verwechslung zwischen allen Klassen
   c) fasse die Klassen $\mathcal{C}_i, \mathcal{C}_j$ zusammen, für die $V(\mathcal{C}_i, \mathcal{C}_j)$ maximal ist.
   d) falls gewünschte Klassenanzahl noch nicht erreicht, gehe zu b)
   e) Ende

Im Gegensatz zu den bisherigen Verfahren, unterscheiden sich hier die verwendeten Viseme für Graustufenbilder und LDA-Bilder, da der Erkenner bei den verschiedenen Eingabebildern unterschiedliche Phoneme miteinander verwechselt.

Anhang D zeigt die Reihenfolge der Zusammenfassungen bei dieser Methode, jeweils für Graustufen- und LDA-Bilder, und entsprechende Einteilungen in 41 Viseme.

5.5 Ergebnisse

Es wurden Ergebnisse mit 31, 41 und 51 Visemen untersucht.

Tabelle 5.1 zeigt die rein visuellen Erkennungsraten, die mit dieser Methode erzielt wurden.

Verfahren	#Viseme	Graustufenb.	LDA
Conf.matrix	31	18.8 %	18.2 %
Conf.matrix	41	19.4 %	39.4 %
Conf.matrix	51	30.9 %	50.9 %
ohne Viseme	(63)	28.5 %	55.2 %
von Hand	42	29.7 %	52.7 %

Tabelle 5.1: visuelle Erkennungsraten mit Konfusionsmatrizen-Ansatz

Es zeigt sich, daß mit einer Anzahl von 51 Visemen sowohl mit Graustufenbildern, als auch mit LDA-Bildern, die besten Ergebnisse erzielt werden. Dabei wird das Referenzergebnis bei Eingabe von Graustufenbilder leicht übertroffen, bei LDA-Bildern hingegen nicht ganz erreicht.

Tabelle 5.2 zeigt die Erkennungsraten bei kombinierter akustischer und visueller Erkennung, mit Graustufenbildern als visueller Eingabe.

Verfahren	#Viseme	unverrauscht	16 db SNR	8 db SNR
Conf.matrix	31	95.8 %	70.9 %	47.9 %
Conf.matrix	41	94.5 %	70.3 %	46.7 %
Conf.matrix	51	93.9 %	70.9 %	47.3 %
ohne Viseme	(63)	96.4 %	70.3 %	50.3 %
von Hand	42	95.8 %	73.9 %	50.3 %
rein akustisch	-	93.9 %	64.8 %	39.4 %

Tabelle 5.2: kombinierte Erkennungsraten mit Konfusionsmatrizen-Ansatz und Eingabe von Graustufenbildern

Im Gegensatz zur rein visuellen Erkennung wird bei der kombinierten Erkennung das beste Ergebnis (mit automatischer Klassenfindung) mit nur 31 Visemen erreicht. Dabei werden die Referenzergebnisse bei verrauschten akustischen Daten knapp verfehlt. Bei unverrauschten akustischen Daten wird auch mit den 31 automatisch erzeugten Visemen ein gleichwertiges Erkennungsergebnis wie mit den von Hand erzeugten Visemen erreicht.

Tabelle 5.3 zeigt die entsprechenden Ergebnisse mit LDA-Bildern.

Verfahren	#Viseme	unverrauscht	16 db SNR	8 db SNR
Conf.matrix	31	93.9 %	73.9 %	55.8 %
Conf.matrix	41	95.2 %	73.3 %	55.8 %
Conf.matrix	51	95.2 %	76.3 %	63.6 %
von Hand	42	96.4 %	77.0 %	63.0 %
rein akustisch	-	93.9 %	64.8 %	39.4 %

Tabelle 5.3: kombinierte Erkennungsraten mit Konfusionsmatrizen-Ansatz und Eingabe von LDA-Bildern

Auch mit Eingabe von LDA-Bildern wurden mit diesem Ansatz bei der kombinierten Erkennung etwa gleichgute Ergebnisse wie bei den von Hand gefundenen Visemen erzielt. Dabei war die Einteilung in 51 Viseme am Besten.

6. Vergleich der vorgestellten Verfahren

In diesem Kapitel sollen die einzelnen Verfahren miteinander verglichen werden. Abbildung 6.1 zeigt die besten Ergebnisse der einzelnen Verfahren bei rein visueller Erkennung.

Verfahren	#Viseme	Word Accuracy
Mittelwertbilder	32	18.2 %
DTW-Verfahren	52	30.9 %
Confusion-Matrix	51	30.9 %
von Hand	42	29.7 %
ohne Viseme	(63)	28.5 %

Tabelle 6.1: beste visuelle Erkennungsraten mit den einzelnen Verfahren

6.1 Verfahren mit Mittelwertbildern

Mit diesem Verfahren wurden sowohl im Vergleich mit den handbestimmten Visemen, als auch im Vergleich mit den anderen hier vorgestellten Verfahren, nur sehr schlechte Ergebnisse erzielt. Als Verfahren zur automatischen Visembestimmung ist es daher nicht geeignet.

6.2 DTW-Verfahren versus Verfahren mit Konfusionsmatrix

Sowohl beim DTW-Verfahren, als auch beim Clustern anhand der Konfusionsmatrix, wurden ähnlich gute Ergebnisse wie bei handbestimmten Visemen erreicht. In manchen Fällen konnte das Erkennungsergebnis sogar leicht verbessert werden. Deshalb könnten beide Verfahren zur automatischen Visemfindung dienen. Allerdings betrug die Laufzeit des DTW-Verfahrens circa eine Woche, gegenüber einer "Laufzeit" des Clusterings anhand der Kon-

fusionsmatrix im Sekundenbereich. Damit ist das letztere Verfahren im Hinblick auf Wartezeiten und begrenzte Rechnerresourcen sehr viel praktikabler.

6.3 Zufällige Visembestimmung

Um feststellen zu können, ob die verwendeten Verfahren sinnvolle Ergebnisse liefern, wurden zum Vergleich, mehrere Visemklassen zufällig bestimmt, und der Erkenner damit trainiert. Tabelle 6.2 zeigt die dabei erzielten Erkennungsergebnisse (Word Accuracy) mit Graustufenbildern.

#Viseme	31	41	51
1. Test	22.4 %	23.0 %	24.9 %
2. Test	24.9 %	24.2 %	24.2 %
Ø	23.3 %	24.0 %	24.6 %

Tabelle 6.2: visuelle Erkennungsraten mit zufällig bestimmten Visemen

Man sieht, daß sowohl die von Hand gefundenen, als auch die mit DTW-Verfahren und Confusion-Matrix-Verfahren gefundenen Viseme, bessere Ergebnisse liefern.

6.4 Erkennung ohne Viseme

Es zeigte sich, daß bei dieser Aufgabe (sprecherabhängige, kontinuierliche Buchstabenerkennung), durch die Verwendung von Visemen nur eine sehr geringe Verbesserung der Erkennungsrate erzielt wird. Es werden hier also – entgegen bisheriger Annahmen – kaum schlechtere Ergebnisse erzielt, als wenn man versucht alle Phoneme zu klassifizieren. Der einzige Vorteil hier bestünde also nur in einer geringeren Anzahl von zu unterscheidenden Klassen.

Es ist allerdings wahrscheinlich, daß die Verwendung von Visemen bei anderen Aufgaben, z. B. bei sprecherunabhängiger Erkennung, eine größere Steigerung der Erkennungsrate mit sich bringen.

7. Zusammenfassung

Es wurde gezeigt, daß es möglich ist, Phoneme automatisch zu Visemen zusammenzufassen, und damit vergleichbare Ergebnisse wie mit handbestimmten Visemen zu erzielen.

Damit ist es in Zukunft möglich, ohne spezielles Expertenwissen über Viseme oder Verwechslungen unterschiedlicher Phoneme, Viseme automatisch, innerhalb kürzester Zeit zu bestimmen.

So lassen sich nun auch beispielsweise beim Wechsel von einem sprecherabhängigen System zu einem sprecherunabhängigen System, leicht neue, möglicherweise unterschiedliche Viseme finden. Dies könnte unter Umständen sogar schon bei Änderungen der Beleuchtungssituation sinnvoll sein, da diese sich möglicherweise auf die Unterscheidbarkeit einzelner Phoneme auswirken könnte.

A. Benutzte Phoneme

A.1 Beschreibung der Buchstaben

Die zu erkennenden Buchstaben des deutschen Alphabetes wurden mit 63 Phonemen wie folgt beschrieben:

a:	/gs/	ahI	ahF					
b:	bI	b-eh	ehF					
c:	tI	s	s-eh	ehF				
d:	dI	d-eh	eh	F				
e:	/gs/	ehI	ehF					
f:	/gs/	aeI	ae-f	fF				
g:	gI	g-eh	ehF					
h:	hI	h-ah	ahF					
i:	/gs/	ieI	ieF					
j:	jI	j-o	o-t	tF				
k:	kI	k-ah	ahF					
l:	/gs/	aeI	ae-l	lF				
m:	/gs/	aeI	ae-m	mF				
n:	/gs/	aeI	ae-n	nF				
o:	/gs/	ohI	ohF					
p:	pI	p-eh	ehF					
q:	kI	k-uh	uhF					
r:	/gs/	aeI	ae-r	rF				
s:	/gs/	aeI	ae-s	s	F			
t:	tI	t-eh	ehF					
u:	/gs/	uhI	uhF					
v:	fI	f-au	auF					
w:	vI	v-eh	ehF					
x:	/gs/	iI	i-k	k-s	sF			
y:	/gs/	ueI	p	s	i	l	o	nF
z:	tI	s-t	t-ae	ae-t	tF			
silence:	si1	si2						

A.2 Phoneme

Zur Beschreibung der Buchstaben wurden also folgende 63 Phoneme verwendet:

si1	si2	/gs/	ahI	ahF	bI	b-eh
ehF	tI	s	s-eh	dI	d-eh	ehI
aeI	ae-f	fF	gI	g-eh	hI	h-ah
ieI	ieF	jI	j-o	o-t	tF	kI
k-ah	ae-l	lF	ae-m	mF	ae-n	nF
ohI	ohF	pI	p-eh	k-uh	uhF	ae-r
rF	ae-s	sF	t-eh	uhI	fI	f-au
auF	vI	v-eh	iI	i-k	k-s	ueI
p	i	l	o	s-t	t-ae	ae-t

Hierbei bezeichnet /gs/ das Phonem für "glottal stop".

A.3 Von Hand bestimmte Viseme

Folgende von Hand bestimmte Viseme werden bisher zur kombinierten Erkennung benutzt:

Visem 1: ae-r
Visem 2: ahF
Visem 3: ahI
Visem 4: bI, mF, p, pI
Visem 5: b-eh, p-eh
Visem 6: kI, nF, tF, dI
Visem 7: k-ah
Visem 8: d-eh, g-eh, t-eh
Visem 9: k-uh
Visem 10: rF
Visem 11: gI
Visem 12: ae-n
Visem 13: ae-f
Visem 14: ae-l
Visem 15: ae-m
Visem 16: ae-s
Visem 17: ae-t
Visem 18: ehF
Visem 19: aeI, ehI
Visem 20: fF, fI, vI
Visem 21: v-eh, f-au
Visem 22: h-ah

Visem 23:	hI
Visem 24:	i-k
Visem 25:	i
Visem 26:	j-o, o-t
Visem 27:	ieF
Visem 28:	iI, ieI, jI
Visem 29:	l, lF
Visem 30:	o
Visem 31:	ohF
Visem 32:	ohI
Visem 33:	k-s, s, sF, tI
Visem 34:	s-eh, s-t, t-ae
Visem 35:	si2
Visem 36:	si1
Visem 37:	auF
Visem 38:	ueI
Visem 39:	uhF
Visem 40:	uhI

B. Klassenbildung mit Mittelwertbildern

B.1 Einzelschritte des Clusterings

Im Folgenden werden die ersten dreißig Schritte des hierarchischen Clusterings mit Mittelwertbildern (siehe Kapitel 3) aufgelistet.

In der ersten Spalte werden die zusammengefaßten Klassen angegeben, wobei diese jeweils durch ihr erstes Element (Phonem) repräsentiert werden. In der nächsten Spalte wird die daraus entstehende neue Klasse angegeben.

```
 2. (ehF,ehI)   : {ehF, ehI}
 3. (aeI,ehF)   : {aeI, ehF, ehI}
 4. (h-ah,hI)   : {h-ah, hI}
 5. (ae-n,nF)   : {ae-n, nF}
 6. (ahF,ahI)   : {ahF, ahI}
 7. (d-eh,dI)   : {d-eh, dI}
 8. (ae-s,sF)   : {ae-s, sF}
 9. (g-eh,gI)   : {g-eh, gI}
10. (ieF,ieI)   : {ieF, ieI}
11. (ae-n,aeI)  : {ae-n, nF, aeI, ehF, ehI}
12. (ae-l,lF)   : {ae-l, lF}
13. (ae-n,d-eh): {ae-n, nF, aeI, ehF, ehI, d-eh, dI}
14. (ae-l,ae-n): {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI}
15. (j-o,jI)    : {j-o, jI}
16. (ae-l,ae-r): {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r}
17. (ae-s,ieF)  : {ae-s, sF, ieF, ieI}
18. (ae-s,tI)   : {ae-s, sF, ieF, ieI, tI}
19. (ae-s,s)    : {ae-s, sF, ieF, ieI, tI, s}
20. (ae-l,ahF)  : {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r, ahF,
                   ahI}
21. (ae-m,mF)   : {ae-m, mF}
22. (ae-l,g-eh): {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r, ahF,
                  ahI, g-eh, gI}
```

```
23. (ae-l,rF)   : {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r, ahF,
                   ahI, g-eh, gI, rF}
24. (j-o,o-t)   : {j-o, jI, o-t}
25. (i-k,iI)    : {i-k, iI}
26. (ohF,uhF)   : {ohF, uhF}
27. (ohF,tF)    : {ohF, uhF, tF}
28. (kI,ohF)    : {kI, ohF, uhF, tF}
29. (ae-l,ae-s): {ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r, ahF,
                   ahI, g-eh, gI, rF, ae-s, sF, ieF, ieI, tI, s}
30. (i-k,k-s)   : {i-k, iI, k-s}
31. (i,l)       : {i, l}
```

B.2 Einteilung in einunddreißig Viseme

Im Folgenden wird exemplarisch eine Einteilung in einunddreißig Viseme gezeigt. Dabei wurde das Phonem/Visem für "Stille" – /si/ – zu den dreißig durch Clustering bestimmten Visemen, hinzugefügt.

```
Visem 0: { si }
Visem 1: { ae-f }
Visem 2: { ae-l, lF, ae-n, nF, aeI, ehF, ehI, d-eh, dI, ae-r, ahF,
           ahI, g-eh, gI, rF, ae-s, sF, ieF, ieI, tI, s }
Visem 3: { ae-m, mF }
Visem 4: { ae-t }
Visem 5: { auF }
Visem 6: { b-eh }
Visem 7: { bI }
Visem 8: { f-au }
Visem 9: { fF }
Visem 10: { fI }
Visem 11: { h-ah, hI }
Visem 12: { i, l }
Visem 13: { i-k, iI, k-s }
Visem 14: { j-o, jI, o-t }
Visem 15: { k-ah }
Visem 16: { k-uh }
Visem 17: { kI, ohF, uhF, tF }
Visem 18: { o }
Visem 19: { ohI }
Visem 20: { p }
Visem 21: { p-eh }
```

Visem 22: { pI }
Visem 23: { s-eh }
Visem 24: { s-t }
Visem 25: { t-ae }
Visem 26: { t-eh }
Visem 27: { ueI }
Visem 28: { uhI }
Visem 29: { v-eh }
Visem 30: { vI }

C. Klassenbildung mit DTW-Verfahren

C.1 Einzelschritte des Clusterings

Im folgenden werden die ersten dreißig Schritte des hierarchischen Clusterings mit Dynamischer zeitlicher Anpassung (siehe Kapitel 4) aufgelistet.

Die zusammengefaßten Klassen stehen in der ersten Spalte und werden wieder durch ihr erstes Element repräsentiert. Danach wird die aus der Zusammenfassung resultierende neue Klasse angegeben.

1. (auF,f-au) : { auF, f-au }
2. (auF,fI) : { auF, f-au, fI }
3. (i-k,k-s) : { i-k, k-s }
4. (i-k,iI) : { i-k, k-s, iI }
5. (auF,i-k) : { auF, f-au, fI, i-k, k-s, iI }
6. (k-uh,uhI) : { k-uh, uhI }
7. (j-o,jI) : { j-o, jI }
8. (k-ah,t-eh): { k-ah, t-eh }
9. (i,l) : { i, l }
10. (auF,k-ah) : { auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh }
11. (i,o) : { i, l, o }
12. (auF,kI) : { auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI }
13. (ae-t,t-ae): { ae-t, t-ae }
14. (j-o,o-t) : { j-o, jI, o-t }
15. (auF,v-eh) : { auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh }
16. (h-ah,hI) : { h-ah, hI }
17. (ae-r,auF) : { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh }
18. (k-uh,ohI) : { k-uh, uhI, ohI }
19. (ae-r,p-eh): { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh }
20. (ae-r,rF) : { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF }

21. (ae-r,b-eh): { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF, b-eh }
22. (ae-t,s-t) : { ae-t, t-ae, s-t }
23. (g-eh,gI) : { g-eh, gI }
24. (i,uhF) : { i, l, o, uhF }
25. (j-o,k-uh) : { j-o, jI, o-t, k-uh, uhI, ohI }
26. (ae-r,i) : { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF, b-eh, i, l, o, uhF }
27. (ae-l,ae-r): { ae-l, ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF, b-eh, i, l, o, uhF }
28. (ae-l,ehF) : { ae-l, ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF, b-eh, i, l, o, uhF, ehF }
29. (ahF,h-ah) : { ahF, h-ah, hI }
30. (ae-l,d-eh): { ae-l, ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF, b-eh, i, l, o, uhF, ehF, d-eh }
31. (ehI,g-eh) : { ehI, g-eh, gI }

C.2 Einteilung in einundvierzig Viseme

Im Folgenden wird exemplarisch eine Einteilung in einundvierzig Viseme gezeigt. Dabei wurde das Phonem/Visem für "Stille" – /si/ – zu den vierzig durch Clustering bestimmten Visemen, hinzugefügt.

Visem 0: { si }
Visem 1: { ae-f }
Visem 2: { ae-l }
Visem 3: { ae-m }
Visem 4: { ae-n }
Visem 5: { ae-r, auF, f-au, fI, i-k, k-s, iI, k-ah, t-eh, kI, v-eh, p-eh, rF }
Visem 6: { ae-s }
Visem 7: { ae-t, t-ae }
Visem 8: { aeI }
Visem 9: { ahF }
Visem 10: { ahI }
Visem 11: { b-eh }
Visem 12: { bI }
Visem 13: { d-eh }
Visem 14: { dI }
Visem 15: { ehF }
Visem 16: { ehI }
Visem 17: { fF }

Visem 18: { g-eh }
Visem 19: { gI }
Visem 20: { h-ah, hI }
Visem 21: { i, l, o }
Visem 22: { ieF }
Visem 23: { ieI }
Visem 24: { j-o, jI, o-t }
Visem 25: { k-uh, uhI, ohI }
Visem 26: { lF }
Visem 27: { mF }
Visem 28: { nF }
Visem 29: { ohF }
Visem 30: { p }
Visem 31: { pI }
Visem 32: { s }
Visem 33: { s-eh }
Visem 34: { s-t }
Visem 35: { sF }
Visem 36: { tF }
Visem 37: { tI }
Visem 38: { ueI }
Visem 39: { uhF }
Visem 40: { vI }

D. Klassenbildung beim Clustering anhand der Konfusionsmatrix

D.1 Clustering mit Graustufenbildern

D.1.1 Einzelschritte des Clusterings

Im folgenden werden die ersten dreißig Schritte des hierarchischen Clusterings anhand der Konfusionsmatrix des Neuronalen Netzes (siehe Kapitel 5), bei Eingabe von Graustufenbildern, aufgelistet.

1. (uhI,k-uh) : { uhI, k-uh }
2. (fI,fF) : { fI, fF }
3. (o-t,j-o) : { o-t, j-o }
4. (ahF,ahI) : { ahF, ahI }
5. (sF,ieF) : { sF, ieF }
6. (sF,ieI) : { sF, ieF, ieI }
7. (sF,s-eh) : { sF, ieF, ieI, s-eh }
8. (sF,ae-s) : { sF, ieF, ieI, s-eh, ae-s }
9. (sF,ehF) : { sF, ieF, ieI, s-eh, ae-s, ehF }
10. (uhF,ohF) : { uhF, ohF }
11. (ae-r,ae-l): { ae-r, ae-l }
12. (ae-r,ahF) : { ae-r, ae-l, ahF, ahI }
13. (uhI,uhF) : { uhI, k-uh, uhF, ohF }
14. (p-eh,b-eh): { p-eh, b-eh }
15. (sF,ae-r) : { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI }
16. (nF,lF) : { nF, lF }
17. (uhI,ohI) : { uhI, k-uh, uhF, ohF, ohI }
18. (f-au,fI) : { f-au, fI, fF }
19. (pI,mF) : { pI, mF }
20. (pI,ae-m) : { pI, mF, ae-m }
21. (sF,ehI) : { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI, ehI }
22. (sF,aeI) : { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI, ehI, aeI }

```
23. (v-eh,vI)   : { v-eh, vI }
24. (p-eh,pI)   : { p-eh, b-eh, pI, mF, ae-m }
25. (sF,nF)     : { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI,
                    ehI, aeI, nF, lF }
26. (auF,o-t)   : { auF, o-t, j-o }
27. (f-au,ae-f): { f-au, fI, fF, ae-f }
28. (sF,p-eh)   : { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI,
                    ehI, aeI, nF, lF, p-eh, b-eh, pI, mF, ae-m }
29. ( uhI,tF)   : { uhI, k-uh, uhF, ohF, ohI, tF }
30. (sF,ae-n)   : {sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI,
                    ehI, aeI, nF, lF, p-eh, b-eh, pI, mF, ae-m, ae-n }
```

D.1.2 Einteilung in einundvierzig Viseme

Im Folgenden wird exemplarisch eine Einteilung in einundvierzig Viseme gezeigt. Dabei wurde das Phonem/Visem für "Stille" – /si/ – zu den vierzig durch Clustering bestimmten Visemen, hinzugefügt.

Visem 0: { si }
Visem 1: { bI }
Visem 2: { tI }
Visem 3: { s }
Visem 4: { dI }
Visem 5: { d-eh }
Visem 6: { ehI }
Visem 7: { aeI }
Visem 8: { ae-f }
Visem 9: { gI }
Visem 10: { g-eh }
Visem 11: { hI }
Visem 12: { h-ah }
Visem 13: { jI }
Visem 14: { o-t, j-o }
Visem 15: { tF }
Visem 16: { kI }
Visem 17: { k-ah }
Visem 18: { ae-n }
Visem 19: { nF, lF }
Visem 20: { pI, mF, ae-m }
Visem 21: { p-eh, b-eh }
Visem 22: { rF }
Visem 23: { sF, ieF, ieI, s-eh, ae-s, ehF, ae-r, ae-l, ahF, ahI }

Visem 24: { t-eh }
Visem 25: { uhI, k-uh, uhF, ohF, ohI }
Visem 26: { f-au, fI, fF }
Visem 27: { auF }
Visem 28: { vI }
Visem 29: { v-eh }
Visem 30: { iI }
Visem 31: { i-k }
Visem 32: { k-s }
Visem 33: { ueI }
Visem 34: { p }
Visem 35: { i }
Visem 36: { l }
Visem 37: { o }
Visem 38: { s-t }
Visem 39: { t-ae }
Visem 40: { ae-t }

D.2 Clustering mit LDA-Bildern

D.2.1 Einzelschritte des Clusterings

Im folgenden werden die ersten dreißig Schritte des hierarchischen Clusterings anhand der Konfusionsmatrix des Neuronalen Netzes (siehe Kapitel 5), bei Eingabe von LDA-Bildern, aufgelistet.

1. (fF,ae-f) : { fF, ae-f }
2. (ieF,ieI) : { ieF, ieI }
3. (l,o-t) : { l, o-t }
4. (auF,j-o) : { auF, j-o }
5. (mF,ae-m) : { mF, ae-m }
6. (ahF,ahI) : { ahF, ahI }
7. (uhI,k-uh) : { uhI, k-uh }
8. (ae-s,s-eh): { ae-s, s-eh }
9. (sF,ae-s) : { sF, ae-s, s-eh }
10. (mF,bI) : { mF, ae-m, bI }
11. (l,i) : { l, o-t, i }
12. (uhI,uhF) : { uhI, k-uh, uhF }
13. (lF,ae-l) : { lF, ae-l }
14. (ehI,ehF) : { ehI, ehF }
15. (v-eh,vI) : { v-eh, vI }
16. (p-eh,g-eh): { p-eh, g-eh }

17. (h-ah,hI) : { h-ah, hI }
18. (ae-n,aeI) : { ae-n, aeI }
19. (ae-n,ehI) : { ae-n, aeI, ehI, ehF }
20. (l,tF) : { l, o-t, i, tF }
21. (k-s,ieF) : { k-s, ieF, ieI }
22. (auF,ohI) : { auF, j-o, ohI }
23. (auF,jI) : { auF, j-o, ohI, jI }
24. (k-s,sF) : { k-s, ieF, ieI, sF, ae-s, s-eh }
25. (k-s,ae-n) : { k-s, ieF, ieI, sF, ae-s, s-eh, ae-n, aeI, ehI, ehF }
26. (f-au,fF) : { f-au, fF, ae-f }
27. (auF,ohF) : { auF, j-o, ohI, jI, ohF }
28. (auF,uhI) : { auF, j-o, ohI, jI, ohF, uhI, k-uh, uhF }
29. (k-s,ahF) : { k-s, ieF, ieI, sF, ae-s, s-eh, ae-n, aeI, ehI, ehF, ahF, ahI }
30. (v-eh,tI) : { v-eh, vI, tI }

D.2.2 Einteilung in einundvierzig Viseme

Im Folgenden wird exemplarisch eine Einteilung in einundvierzig Viseme gezeigt. Dabei wurde das Phonem/Visem für "Stille" – /si/ – zu den vierzig durch Clustering bestimmten Visemen, hinzugefügt.

Visem 0: { si }
Visem 1: { ahF, ahI }
Visem 2: { b-eh }
Visem 3: { tI }
Visem 4: { s }
Visem 5: { dI }
Visem 6: { d-eh }
Visem 7: { fF, ae-f }
Visem 8: { gI }
Visem 9: { h-ah, hI }
Visem 10: { ieF, ieI }
Visem 11: { jI }
Visem 12: { kI }
Visem 13: { k-ah }
Visem 14: { lF ae-l }
Visem 15: { mF, ae-m, bI }
Visem 16: { ae-n, aeI, ehI, ehF }
Visem 17: { nF }
Visem 18: { ohI }
Visem 19: { ohF }
Visem 20: { pI }

Visem 21: { p-eh, g-eh }
Visem 22: { ae-r }
Visem 23: { rF }
Visem 24: { sF, ae-s, s-eh }
Visem 25: { t-eh }
Visem 26: { uhI, k-uh, uhF }
Visem 27: { fI }
Visem 28: { f-au }
Visem 29: { auF, j-o }
Visem 30: { v-eh, vI }
Visem 31: { iI }
Visem 32: { i-k }
Visem 33: { k-s }
Visem 34: { ueI }
Visem 35: { p }
Visem 36: { l, o-t, i, tF }
Visem 37: { o }
Visem 38: { s-t }
Visem 39: { t-ae }
Visem 40: { ae-t }

Literaturverzeichnis

[1] Christoph Bregler, Hermann Hild, Stefan Manke und Alex Waibel: *Improving connected letter recognition by lipreading*

[2] Christoph Bregler, Stefan Manke, Hermann Hild, Alex Waibel: *Bimodal sensor integration on the example of "speech-reading"*

[3] persönliche Kommunikation mit Chrisoph Bregler, Februar 1995.

[4] P. Duchnowski, U. Meier, A. Waibel: *See Me, Hear me: Integrating Automatic Speech Recognition and Lip-Reading.* International Conference on Spoken Language Processing, ICSLP, 1994.

[5] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, A. Waibel: *Toward Movement-Invariant Automatic lip-reading and speech recognition*, Proc. ICASSP , 1995.

[6] Cletus G. Fisher: *Confusion among visually perceived consonants* in Journal of Speech and Hearing Research, Vol. 11, p. 796-804,1968.

[7] K. Fukunaga: *Introduction to satistical pattern recognition* Boston, Academic Press, 1990

[8] A.J. Goldschen: *Continuous Automatic Speech Recognition by Lipreading.* Dissertation, The School of Engineering and Applied Science of The George Washington University, September 1993.

[9] Hermann Hild, Alex Waibel: *Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network,* EUROSPEECH 93, Berlin, Germany, September 1993, Volume 2, pp. 1481-1484.

[10] Hermann Hild, Alex Waibel: *Connected Letter Recognition with a Multi-State Time Delay Neural Network* NIPS 5, San Marino, CA:Morgan Kaufmann Publishers, 1993.

[11] Pamela L. Jackson: *The theoretical minimal unit for visual speech perception: Visemes and Coarticulation.* in The Volta Review, 90 (5), 1988.

[12] J. Jeffers, M. Barley: *Speechreading* Springfield, IL: Charles C. Thomas, 1971.

[13] Uwe Meier:*Robuste Systemarchitekturen für maschinelles Lippenlesen,* Diplomarbeit, Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany, 1995.

[14] Elmer Owens, Barbara Blazek: *Visemes observed by hearing-impaired and normal-hearing adult viewers.* in Journal of Speech and Hearing Research, Vol. 28, p. 381-393, 1985.

[15] E.D. Petajan: *Automatic Lipreading to enhance speech recognition.* Proc. IEEE Communications Society Global Telecommunications Conference, 1984.

[16] Hiroaki Sakoe, Seibi Chiba: *Dynamic Programming AlgorithmOptimazation for Spoken Word Recognition.* In [19], p. 159 - 163.

[17] P.L. Silsbee und A.C. Bovic: *Audio-visual speech recognition for a vowel discrimination task.* SPIE, 2049:84 95.

[18] D.G.Stork, G. Wolff und E. Levine: *Neural network lipreading system for improved speech recognition.* IJCNN, Juni 1992.

[19] *Readings in speech recognition,* ed. by Waibel, Lee

[20] Brian E. Walden, Robert A. Prosek, Allen A. Montgomery, Charlene K. Scherr und Carla J. Jones: *Effects of training on the visual recognition of consonants* in Journal of Speech and Hearing Research, Vol. 20, p. 130-145, 1977.

[21] Virginia D. Wozniak, Pamela L. Jackson: *Visual vowel and diphtong perception from two horizontal viewing angles* in Journal of Speech and Hearing Research, Vol. 22, p. 354-365, 1979.

[22] B.P. Yuhas, M.H. Goldstein und T.J. Sejnowski: *Integration of acoustic and visual speech signals using neural networks.* IEEE Communications Magazine, p. 65 - 71, November 1989.