Ergänzungen zur Studienarbeit

Adaptive bimodale Sensorfusion für automatische Spracherkennung und Lippenlesen

Im folgenden sind einige Ergebnisse aufgeführt, die auf obiger Studienarbeit aufbauen und im Rahmen einer Hiwi-Tätigkeit erarbeitet wurden.

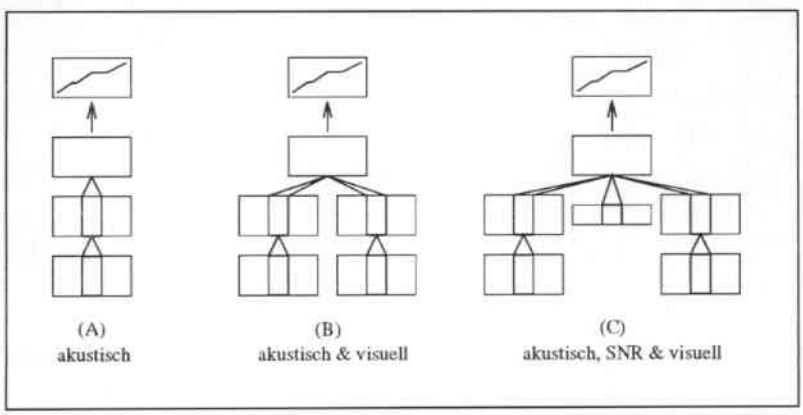
Wolfgang Hürst
Juni 1995

Inhalt

Nachfolgend sind einige Ergebnisse aufgelistet, die bei der Kombination von akustischen und visuellen Eingabesignalen mit verschiedenen Netzarchitekturen (Kombination auf Hidden-Ebene, siehe unten) erzielt wurden. Für eine Beschreibung des Erkenners und der genauen Problemstellung sei auf *Wolfgang Hürst: Adaptive bimodale Sensorfusion für automatische Spracherkennung und Lippenlesen, Studienarbeit, Universität Karlsruhe, Mai 1995* verwiesen.

- In **Teil I** werden die Ergebnisse verschiedener Netzarchitekturen, die jeweils mit der gleichen Trainingsdatenmenge trainiert wurden, miteinander verglichen.
- In **Teil II** sind die Ergebnisse, die mit jeweils einer Netzarchitektur und unterschiedlichen Trainingsmengen erzielt wurden, vergleichend dargestellt.

verwendete Netzarchitekturen



- **Netz (A):** akustische Erkennung; das Netz erhält als Eingabe nur das reine Sprachsignal
- **Netz (B):** kombinierte Erkennung; das Netz erhält das akustische, sowie das „visuelle Sprachsignal" (Bilder der entsprechenden Lippenbewegungen) als Eingabe
- **Netz (C):** kombinierte Erkennung mit zusätzlicher Information; das Netz erhält das akustische Sprachsignal, Bilder der entsprechenden Lippenbewegungen und die aus den akustischen Daten berechneten SNR-Werte (SNR = Signal-to-Noise Ratio) als Eingabe

Datenbasis

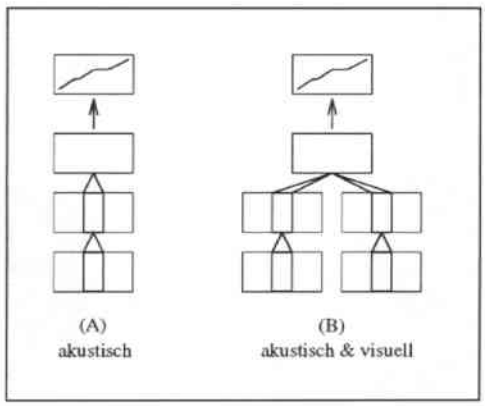
200 unter optimalen Bedingungen aufgenommene Sequenzen eines männlichen Sprechers (mum1&2), davon 170 Sequenzen für das Training und 30 Sequenzen zum Testen der jeweiligen Netze.
Zu Test- und Trainingszwecken wurden diese Sequenzen durch Aufaddieren eines Störgeräuschs künstlich verrauscht, so daß insgesamt folgende Test- und Trainingsdatenbasis zu Verfügung stand:

- CLEAN: unter optimalen Bedingungen aufgenommen; SNR $\approx$ 30 dB)
- NOISE 1 und NOISE 2: konstantes weißes Rauschen; SNR $\approx$ 16 bzw. $\approx$ 8 dB
- INCREASE 1 und INCREASE 2: ansteigendes weißes Rauschen; SNR $\approx$ 30 bis 16 dB bzw. $\approx$ 30 bis 8 dB
- RADIO 1 und RADIO 2: laufendes Radio (Musik); SNR schwankend um $\approx$ 20 bzw. $\approx$ 17 dB
- MOTOR 1 und MOTOR 2: Motor einer aktiven Kamerasteuerung; SNR schwankend um $\approx$ 25 bzw. $\approx$ 11 dB

I.

Ergebnisse der verschiedenen Netzarchitekturen bei jeweils gleicher Trainingsdatenmenge

Netzarchitekturen



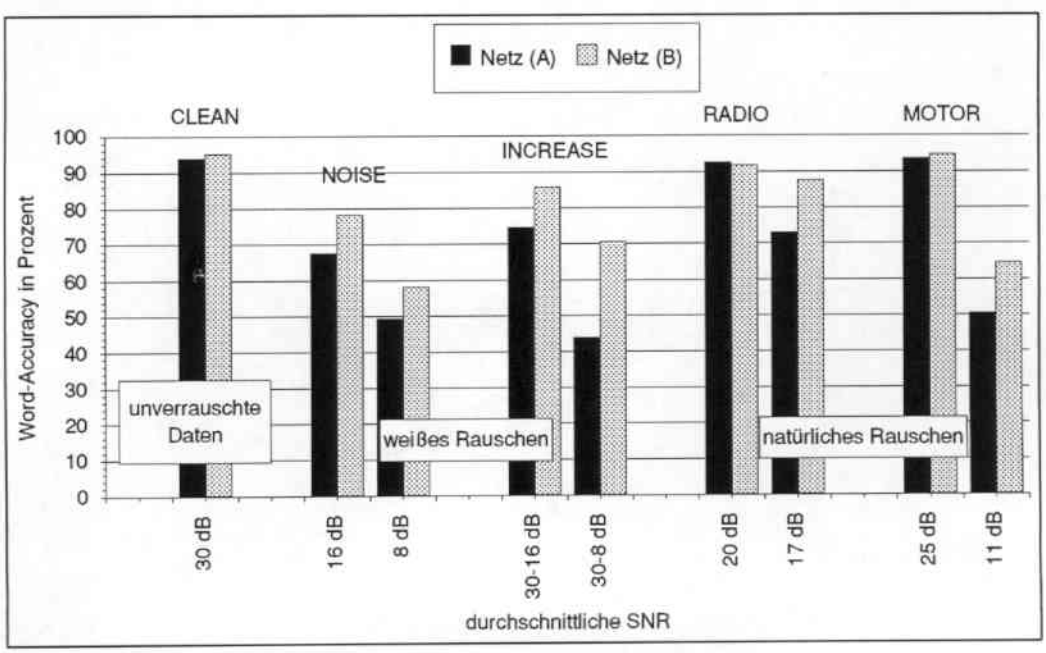
Trainingsmenge

CLEAN (ca. 30 dB SNR)	170 Samples
insgesamt	170 Samples

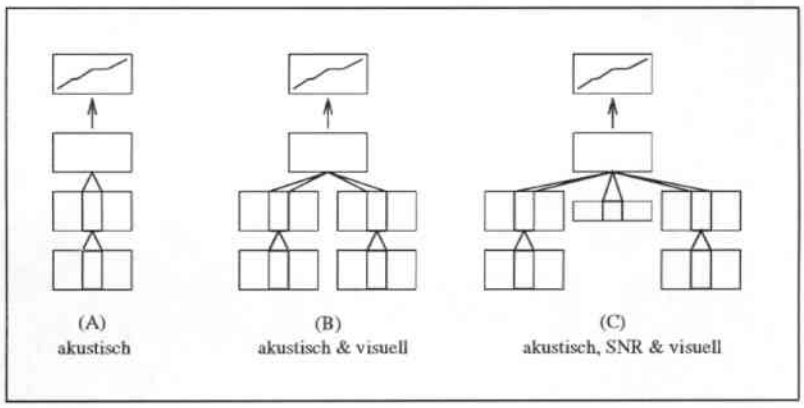
Testergebnisse

Netzarchitektur	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(A)	94.1	67.6	49.4	74.7	44.1	92.4	72.9	93.5	50.6
(B)	95.3	78.2	58.2	85.9	70.6	91.8	87.6	94.7	64.7

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



Netzarchitekturen



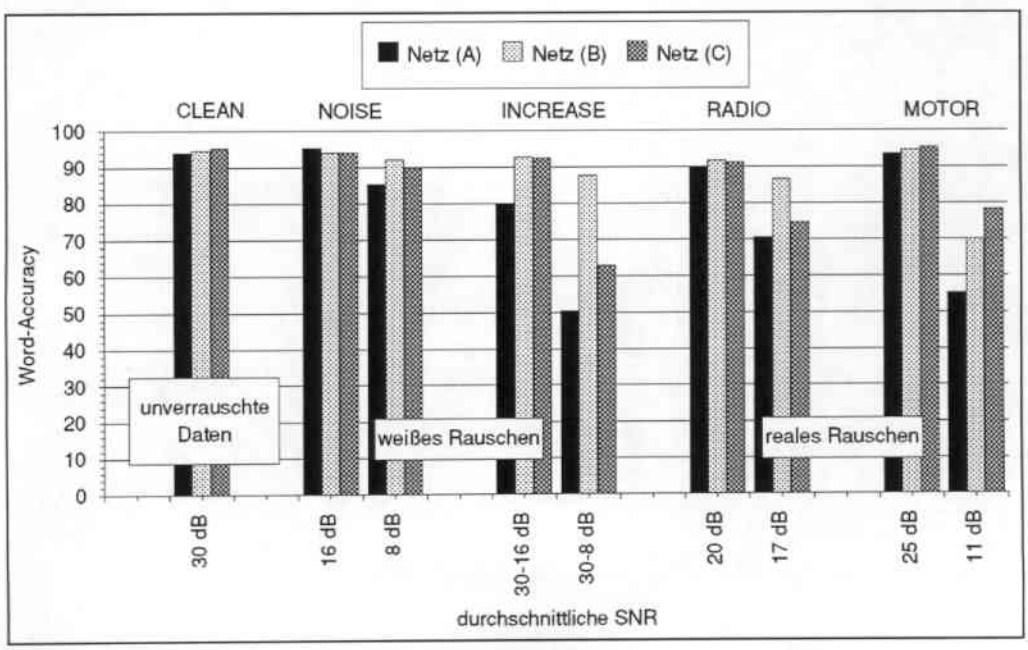
Trainingsmenge

CLEAN (ca. 30 dB SNR)	170 Samples
NOISE 1 (ca. 16 dB SNR)	170 Samples
insgesamt	340 Samples

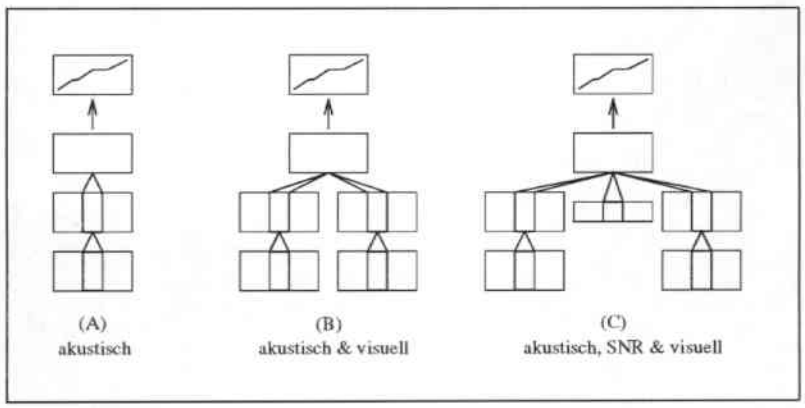
Testergebnisse

Netzarchitektur	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(A)	94.1	95.3	85.3	80.0	50.6	90.0	70.6	93.5	55.3
(B)	94.7	94.1	92.1	92.9	87.6	91.8	86.5	94.7	70.0
(C)	95.3	94.1	90.0	92.4	62.9	91.2	74.7	95.3	78.2

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



Netzarchitekturen



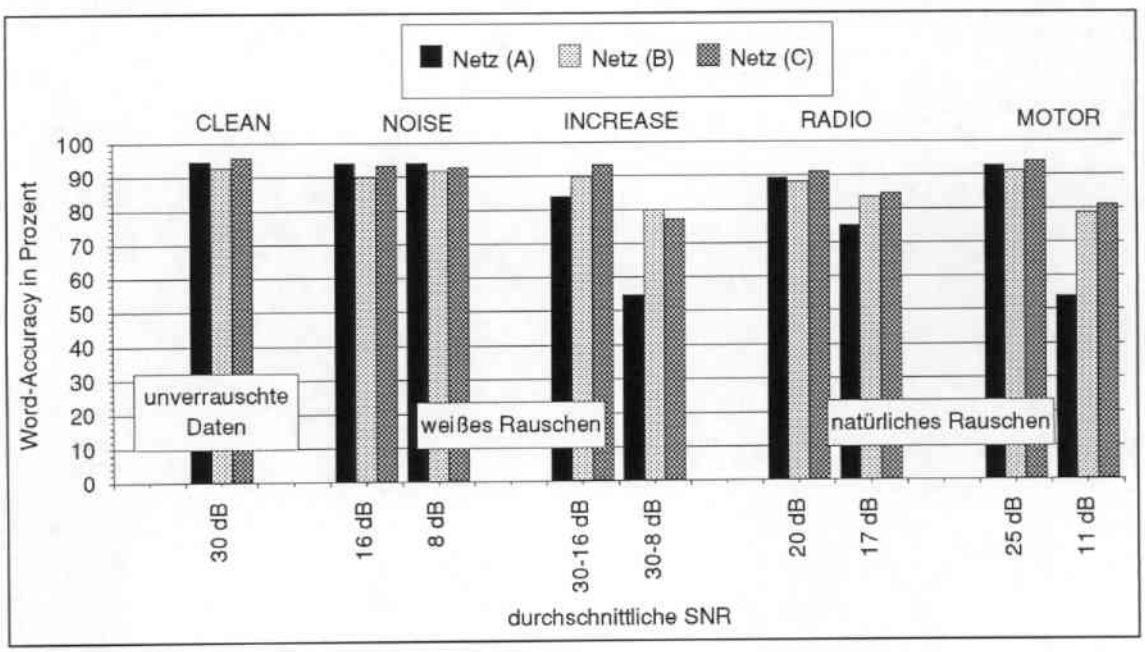
Trainingsmenge

CLEAN (ca. 30 dB SNR)	170 Samples
NOISE 2 (ca. 8 dB SNR)	170 Samples
insgesamt	340 Samples

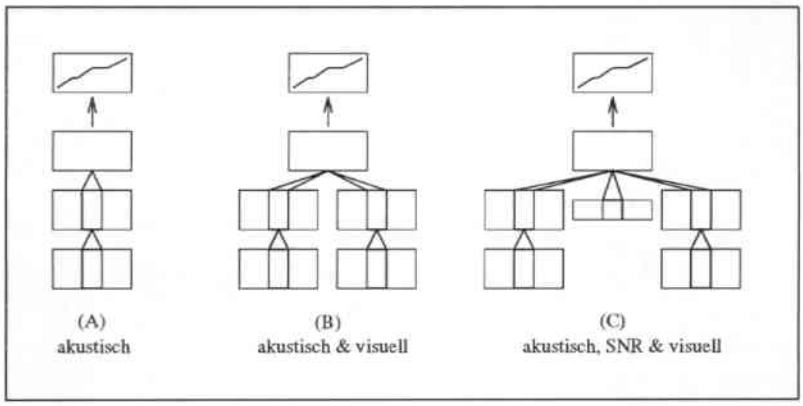
Testergebnisse

Netzarchitektur	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(A)	94.7	94.1	94.1	84.1	54.7	89.4	75.3	92.9	54.1
(B)	92.9	90.0	91.8	90.0	80.0	88.2	83.5	91.2	78.8
(C)	95.9	93.5	92.9	93.5	77.1	91.2	84.7	94.1	81.2

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



Netzarchitekturen



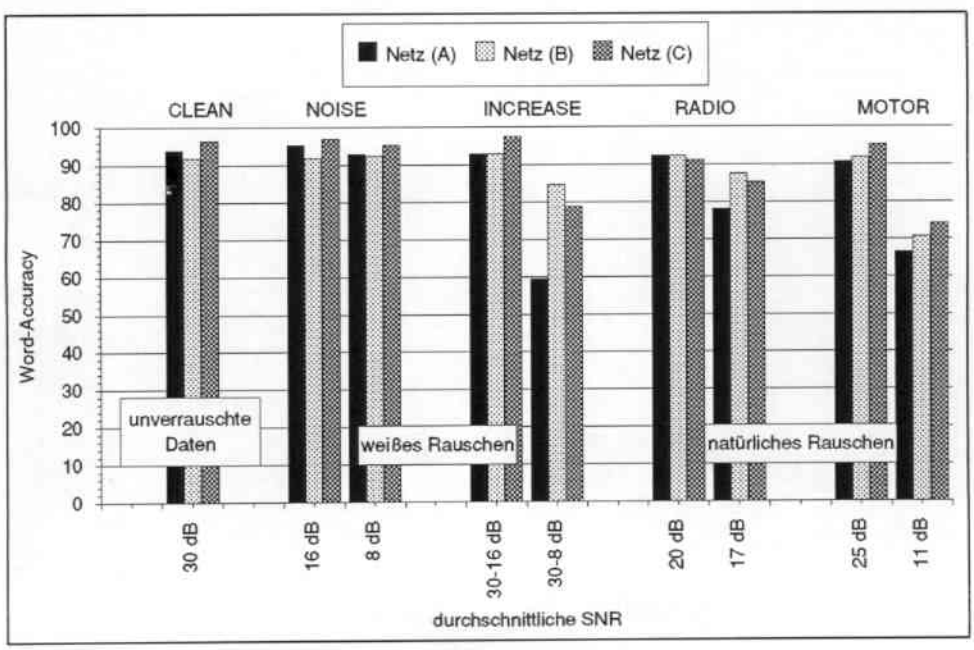
Trainingsmenge

CLEAN (ca. 30 dB SNR)	170 Samples
NOISE 1 (ca. 16 dB SNR)	170 Samples
NOISE 2 (ca. 8 dB SNR)	170 Samples
insgesamt	510 Samples

Testergebnisse

Netzarchitektur	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(A)	94.1	95.3	92.9	92.9	59.4	92.4	78.2	90.6	66.5
(B)	91.8	91.8	92.4	92.9	84.7	92.4	87.6	91.8	70.6
(C)	96.5	97.1	95.3	97.6	78.8	91.2	85.3	95.3	74.1

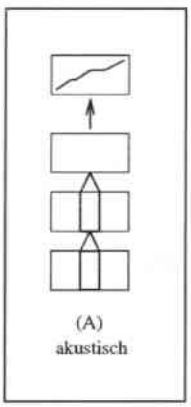
Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



II.

Vergleich der Ergebnisse jeweils einer Netzarchitektur bei Verwendung unterschiedlicher Trainingsdatenmengen

Netzarchitektur



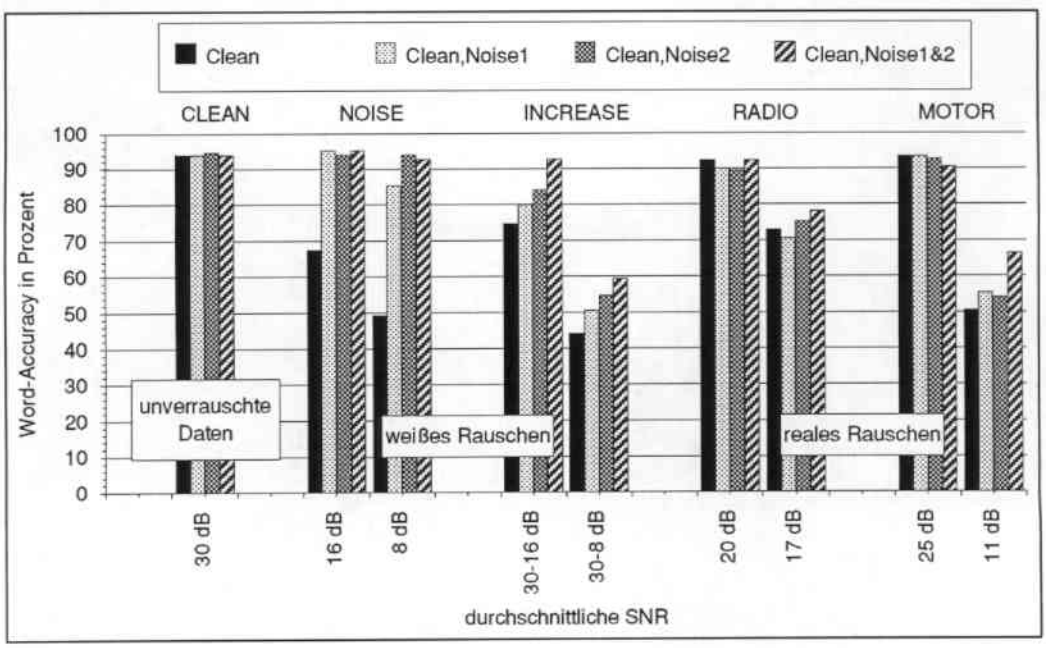
Trainingsmengen

	Menge (1)	Menge (2)	Menge (3)	Menge (4)
CLEAN (ca. 30 dB SNR)	170 Samples	170 Samples	170 Samples	170 Samples
NOISE 1 (ca. 16 dB SNR)		170 Samples		170 Samples
NOISE 2 (ca. 8 dB SNR)			170 Samples	170 Samples
insgesamt	170 Samples	170 Samples	170 Samples	170 Samples

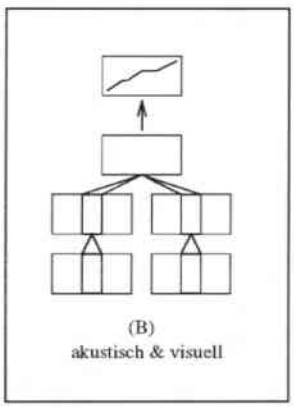
Testergebnisse

Trainings-menge	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(1)	94.1	67.6	49.4	74.7	44.1	92.4	72.9	93.5	50.6
(2)	94.1	95.3	85.3	80.0	50.6	90.0	70.6	93.5	55.3
(3)	94.7	94.1	94.1	84.1	54.7	89.4	75.3	92.9	54.1
(4)	94.1	95.3	92.9	92.9	59.4	92.4	78.2	90.6	66.5

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



Netzarchitektur



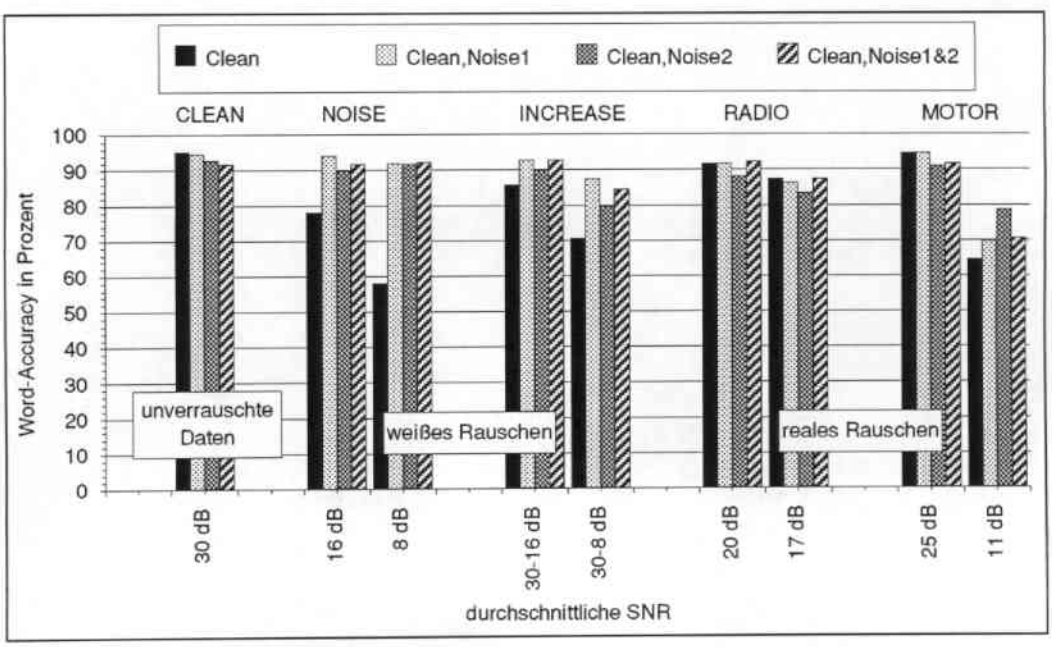
Trainingsmengen

	Menge (1)	Menge (2)	Menge (3)	Menge (4)
CLEAN (ca. 30 dB SNR)	170 Samples	170 Samples	170 Samples	170 Samples
NOISE 1 (ca. 16 dB SNR)		170 Samples		170 Samples
NOISE 2 (ca. 8 dB SNR)			170 Samples	170 Samples
insgesamt	170 Samples	170 Samples	170 Samples	170 Samples

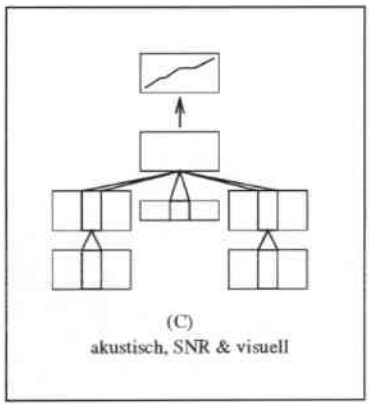
Testergebnisse

Trainings-menge	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(1)	95.3	78.2	58.2	85.9	70.6	91.8	87.6	94.7	64.7
(2)	94.7	94.1	92.1	92.9	87.6	91.8	86.5	94.7	70.0
(3)	92.9	90.0	91.8	90.0	80.0	88.2	83.5	91.2	78.8
(4)	91.8	91.8	92.4	92.9	84.7	92.4	87.6	91.8	70.6

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)



Netzarchitektur



Trainingsmengen

	Menge (1)	Menge (2)	Menge (3)	Menge (4)
CLEAN (ca. 30 dB SNR)	170 Samples	170 Samples	170 Samples	170 Samples
NOISE 1 (ca. 16 dB SNR)		170 Samples		170 Samples
NOISE 2 (ca. 8 dB SNR)			170 Samples	170 Samples
insgesamt	170 Samples	170 Samples	170 Samples	170 Samples

Testergebnisse

Trainings-menge	CLEAN	NOISE 1	NOISE 2	INCR. 1	INCR. 2	RADIO 1	RADIO 2	MOTOR 1	MOTOR 2
(2)	95.3	94.1	90.0	92.4	62.9	91.2	74.7	95.3	78.2
(3)	95.9	93.5	92.9	93.5	77.1	91.2	84.7	94.1	81.2
(4)	96.5	97.1	95.3	97.6	78.8	91.2	85.3	95.3	74.1

Erkennungsraten auf den einzelnen Testdatensets (angegeben ist jeweils die Word-Accuracy in Prozent)