Blinde Quellentrennung als Vorverarbeitung zur robusten Spracherkennung

Jörn Anemüller, Michael Kleinschmidt und Birger Kollmeier
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D–26111 Oldenburg
ane@medi.physik.uni–oldenburg.de http://medi.uni–oldenburg.de/members/ane

I. Einleitung

In diesem Beitrag evaluieren wir den Nutzen blinder Quellentrennung als Vorverarbeitungsstufe zum Zwecke robuster automatischer Spracherkennung. Blinde Quellentrennung (QT) ist eine Signalverarbeitungstechnik, die es ermöglicht, aus mehreren Aufnahmen akustischer Überlagerungen (etwa Sprache im Störgeräusch) die zugrunde liegenden Quellsignale (Sprache getrennt vom Störgeräusch) zu rekonstruieren. Ein spezieller Algorithmus für QT in verhallter Umgebung ist bereits vorgestellt worden [ane99]. Eine potentielle Anwendung solcher Algorithmen besteht in der Störgeräuschbefreiung für die robuste automatische Spracherkennung. Das Perzeptionmodell (PEMO) nach Dau et al. [dau96] wurde bereits zur Merkmalsextraktion in der automatischen Spracherkennung verwendet. Insbesondere in Kombination mit Neuronalen Netzen hat diese gehörgerechte Vorverarbeitung zu einer robusten Erkennungsleistung im Störgeräusch geführt [tch99]. Wir kombinierten den QT–Algorithmus mit einem Einzelworterkennungssystem auf Basis des PEMO, um eine weitere Verbesserung der Erkennungsleistung zu erreichen. Zur Evaluation vergleichen wir die Erkennungsraten bei QT–Vorverarbeitung mit denen ohne Vorverarbeitung und mit alternativen Störgeräuschunterdrückungssystemen. Berücksichtigt werden hierbei Aufnahmesituationen in verhallter und unverhallter Umgebung und bei unterschiedlichen Signal–Rausch Abständen.

II. Blinde Quellentrennung

Algorithmen zur blinden Quellentrennung zeichnen sich dadurch aus, dass sie sehr geringe Annahmen über die vorliegenden Signale machen. Es wird nur vorausgesetzt, dass die Signalquellen voneinander unabhängig sind, und dass die gleiche Anzahl Mikrofone wie Signalquellen vorhanden ist. Insbesondere sind die räumlichen Positionen von Quellen und Mikrofonen unbekannt — daher "blind" —, was blinde Quellentrennung für robuste Spracherkennung besonders interessant macht. Sind die Annahmen erfüllt, dann ist durch Filtern und Überlagern der Mikrofonsignale eine Rekonstruktion der getrennten Quellsignale, bis auf eine prinzipiell unbestimmbare Verzerrung, möglich.
Da die dazu benötigten Filter unbekannt sind, werden sie durch einen Optimierungsalgorithmus iterativ geschätzt, siehe Abb. 2. Die Schlüsselfrage hierzu lautet, wie der Algorithmus bestimmt, ob die rekonstruierten Signale unabhängig oder noch vermischt sind. Kriterien hierfür können aufgrund verschiedener statistischer Maße definiert werden, siehe etwa [nad97]. Der von uns verwendete Algorithmus [ane99] benutzt — motiviert durch Eigenschaften von Sprache — die in verschiedenen Frequenzbändern korrelierte Amplitudenmodulation der Quellsignale. Dazu werden zwischen den rekonstruierten Signalen die Korrelationen der frequenzspezifischen Einhüllenden *frequenzübergreifend* berechnet, also zwischen allen Frequenzbändern f_i des ersten rekonstruierten Signals und allen Frequenzbändern f_j des zweiten rekonstruierten Signals. Die Signale sind dann getrennt, wenn in diesem Sinn eine maximale *De*korrelation erreicht ist. Für eine genauere Beschreibung verweisen wir auf [ane99].
Der benutzte Algorithmus rekonstruiert jeweils die von einer Quelle an den Mikrofonen hervorgerufenen Signale; eine Entfaltung der Raumübertragungsfunktion wird also nicht vorgenommen. Tests mit verschiedenen Signalen zeigen, dass der Algorithmus eine gute Signaltrennung erreicht. Audio–Beispiele sind von der oben genannten WWW–Seite abrufbar.

III. Robuste Spracherkennung

Das Perzeptionsmodell (PEMO) nach Dau et al. [dau96] ist ein funktionelles Modell der Signalverarbeitung im peripheren auditorischen System. Es ist in der Lage, das Antwortverhalten von Versuchspersonen in einer Vielzahl von psychoakustischen Experimenten quantitativ nachzubilden. Das PEMO extrahiert

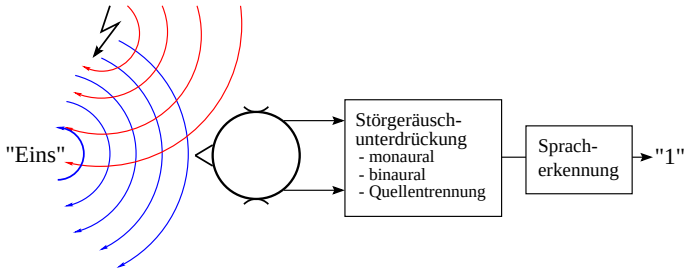

Abb. 1: *Schematische Darstellung des Versuchsaufbaus.*

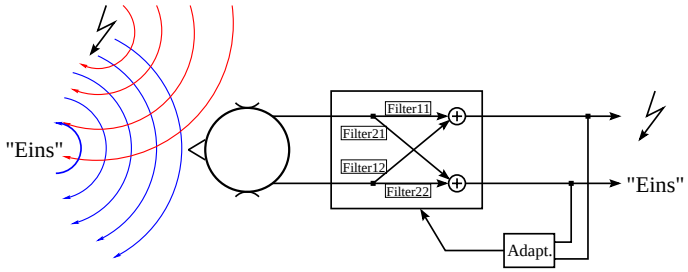

Abb. 2: *Die Architektur des verwendeten Quellentrenners.*

aus einem eintreffenden akustischen Signal die dazugehörige *interne Repräsentation*, welche sich bereits als ein robustes Merkmal für die automatische Spracherkennung bewährt hat [tch99]. Insbesondere in Kombination mit dem lokal–rekurrenten neuronalen Netz (LRNN) als Klassifikator übertrifft die PEMO Vorverarbeitung konventionelle Mel–Cepstralkoeffizienten deutlich an Robustheit gegenüber additiven Störgeräuschen [kas97]. Weiterhin wurde gezeigt, dass sich durch eine Filterung der eintreffenden Zeitsignale mittels monauraler [kle98b] und binauraler [kle98a] Algorithmen zu Störgeräuschreduktion die Erkennungsleistung des PEMO/LRNN Systems bei additiven Störgeräuschen beträchtlich steigern lässt. Vorraussetzung ist dabei allerdings eine zuverlässige Sprachpausendetektion und Stationarität des Störgeräusches für die monaurale, bzw. die Kenntnis der Lage der Schallquellen im Raum für die binaurale Störgeräuschreduktion.

IV. Methoden

Es wurden Kunstkopfaufnahmen von Sprache und Störgeräusch aus reflexionsarmer und aus verhallter Umgebung benutzt. Die Aufnahme in verhallter Umgebung fand in einem Seminarraum mit einer Nachhallzeit T_{60} von ca. 0.5s statt. In allen Fällen betrug der Abstand zwischen Kunstkopf und Lautsprechern etwa 2.5m. Das Sprachsignal kam von vorn, das Störgeräusch von 30 Grad schräg rechts. Diese Signale wurden nachträglich abgemischt bei Signal–Rausch–Abständen (SNR) von -10dB, 0dB und 10dB.
Als Sprachsignale wurden die Wörter "Null" bis "Neun" aus dem ZIFKOM Datensatz verwendet. Insgesamt standen 2000 Artikulationen der Wörter, gesprochen von 200 verschiedenen Sprecherinnen und Sprechern, zur Verfügung. Diese wurden jeweils zur Hälfte als Trainings– und als Testdatensatz für die sprecherunabhängige Spracherkennung benutzt. Als Störgeräusch diente ein sprachähnliches Rauschen ('babble–noise'), das aus der Überlagerung mehrerer Sprachsignale besteht.
Zur Schätzung der optimalen Filter standen dem Quellentrenner für jede Versuchssituation nur die Wörter "Null" bis "Fünf" eines einzigen Sprechers, überlagert mit dem Störgeräusch, zur Verfügung, da die Benutzung des gesamte Testmaterials zu rechenaufwendig gewesen wäre. Die so gefundenen Filter dienten zur Trennung des gesamten Testmaterials in Sprache und Störgeräusch. Die Klassifikation in Sprach– bzw. Störsignal wurde anhand der erzielten Erkennungsrate vorgenommen. Die verwendeten Filter hatten eine Länge von 1536 taps bei einer Samplingrate von 16kHz. Diese große Filterlänge wurde gewählt, um

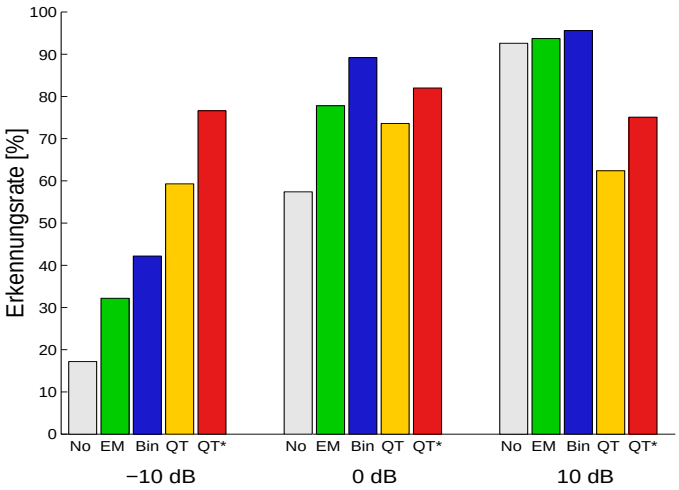

Abb. 3: *Erkennungsleistung im reflexionsarmen Raum für drei SNR–Werte.* No*: Keine Störgeräuschunterdrückung,* EM*: monaural nach Ephraim–Malah,* Bin*: binaurales Richtungsfilter,* QT*: Quellentrenner und reflexionsarum trainiertes LRNN,* QT**: Quellentrenner und verhallt trainiertes LRNN*

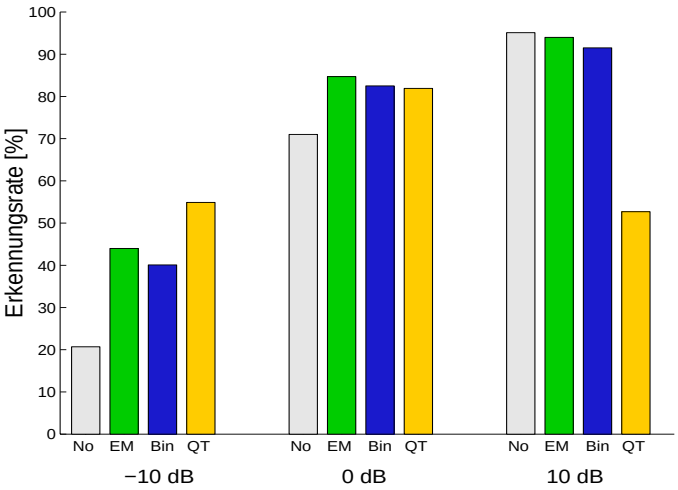

Abb. 4: *Erkennungsleistung verhallter Umgebung. Bezeichnungen wie in Abb. 3, außer* QT*: Quellentrenner und verhallt trainiertes LRNN*

sicherzustellen, dass die Trennung der Signale bei der gegebenen Raumakustik mit langer Nachhallzeit und großem Abstand zwischen den Kunstkopfmikrofonen und den Lautsprechern überhaupt möglich ist.

Zur Spracherkennung wurde die beschriebene Kombination aus PEMO–Vorverarbeitung und LRNN–Klassifikation benutzt. Hierbei wurden zwei neuronale Netzwerke benutzt, die sich darin unterscheiden, dass eines auf reflexionsarm aufgenommenes Trainingsmaterial und das zweite auf verhalltes Trainingsmaterial trainiert wurde.

Der beschriebene Aufbau, siehe Abb. 1, entspricht genau dem der Experimente von Kleinschmidt et al. [kle99], so dass die durch den QT Algorithmus erreichte Verbesserung der Erkennungsleistung direkt mit den bereits vorliegenden Werten für den Ephraim–Malah Algorithmus und das binaurale Richtungsfilter nach Wittkop verglichen werden kann.

V. Ergebnisse

Die Erkennungsraten in reflexionsarmer Umgebung für die drei verwendeten SNR–Werte sind in Abb. 3 dargestellt. Für den SNR von −10dB liegt die Erkennungsrate ohne Vorverabeitung nur unwesentlich über dem Zufallsniveau von 10%. Blinde Quellentrennung erreicht hier eine drastische Verbesserung bis hin zu fast 80% Erkennungsrate. Diese Verbesserung ist signifikant größer als die durch die alternativen Störgeräuschunterdrücker erreichten. Bei 0dB SNR erzielt die Quellentrennung im Vergleich zu den anderen Algorithmen eine vergleichbare bzw. geringfügig niedrigere, jedoch signifikante Verbesserung. Bei einem Pegel von 10dB SNR schließlich bricht die Erkennungsrate bei Quellentrennung ein und liegt sowohl unter der Erkennungsrate ohne Störgeräuschunterdrückung als auch unter dem für 0dB SNR mit Quellentrennung erreichten Wert.

Es fällt auf, dass der auf verhallte Sprache trainierte LRNN–Klassifikator bei Quellentrennung in reflexionsarmer Umgebung besser klassifiziert als der auf reflexionsarm aufgenommene Sprache trainierte LRNN–Klassifikator. Dies ist vermutlich die Folge eines geringfügigen Kammfiltereffektes, der in diesem Fall bei der Quellentrennung als Artefakt auftrat und auch bei Hörtests wahrnehmbar war.

Die Erkennungsraten in verhallter Umgebung sind in Abb. 4 dargestellt. Die Ergebnisse sind vergleichbar mit denen in reflexionsarmer Umgebung: bei −10dB SNR erreicht Quellentrennung die größte Verbesserung aller betrachteten Störgeräuschunterdrücker; bei 0dB ist die Erkennungsrate für alle Störgeräuschunterdrücker ähnlich; bei 10dB bricht die Erkennungsrate bei Quellentrennung ein.

Der Grund für die schlechten Ergebnisse mit Quellentrennung bei 10dB SNR liegt vermutlich darin, dass bei diesem Pegel die Annahmen des Quellentrenners verletzt sind. Eine Schätzung des diffusen Aufnahmerauschens in den Sprachsignalen ergibt, dass dessen Pegel frequenzabhängig im Bereich von etwa −35dB bis −10dB relativ zum Sprachsignal liegt. Bei 10dB SNR erreicht damit das diffuse Aufnahmerauschen in einigen Frequenzbereichen vergleichbare Pegel wie das lokalisierte Störgeräusch. Es stellt damit effektiv eine dritte Signalquelle dar, was die Annahme von nur zwei Signalquellen verletzt, so dass der Quellentrenner keine Signaltrennung mehr erreichen kann.

V. Zusammenfassung

Wegen ihrer minimalen Annahmen über Sprach– und Störsignal ist blinde Quellentrennung interessant als Störgeräuschunterdrückung für robuste Spracherkennung. Der verwendete Quellentrennungsalgorithmus erreicht erfahrungsgemäß eine gute Signaltrennung. Dies resultiert für SNR–Werte von −10dB in einer deutlichen Verbesserung der Erkennungsleistung des Spracherkenners. Bei 0dB SNR ist die Verbesserung durch den Quellentrenner vergleichbar mit den durch alternative Störgeräuschunterdrücker erreichten. Sind jedoch die Annahmen des Quellentrenners verletzt, in diesem Fall durch Aufnahmerauschen bei 10dB SNR, dann kann die Erkennungsleistung zusammenbrechen. Ein weiteres Problem für automatische Spracherkenner können durch die Quellentrennung erzeugte spektrale Veränderungen der Signale, wie etwa Nachhall, darstellen.

Bedanken möchten wir uns bei Klaus Kasper und Herbert Reininger von der Universität Frankfurt dafür, dass sie uns ihre LRNN Implementation zur Benutzung überlassen haben.
Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft im Rahmen des Graduiertenkollegs Psychoakustik unterstützt.

Literatur

[ane99] J. Anemüller: *Correlated modulation: a criterion for blind source separation*, Joint meeting Acoust. Soc. Am. and Europ. Acoust. Assoc., Berlin, Germany, 1999.

[dau96] T. Dau, D. Püschel, A. Kohlrausch: *A quantitive model of the "effective" signal processing in the auditory system I.*, J. Acoust. Soc. Am., 99 (6), pp. 3615–3622, 1996.

[kas97] K.Kasper, H. Reininger, D. Wolf: *Exploiting the Potential of Auditory Preprocessing for Robust Speech Recognition by LRNN*, Proc. ICASSP, vol. 2, pp. 1223–1227, 1997.

[kle98a] M. Kleinschmidt, J. Tchorz, T. Wittkop, V. Hohmann, B. Kollmeier: *Robuste Spracherkennung durch binaurale Richtungsfilterung und gehörgerechte Vorverarbeitung*, Fortschritte der Akustik – DAGA, pp. 396–397, 1998.

[kle98b] M. Kleinschmidt, M. Marzinzik, B. Kollmeier: *Combining Monaural Noise Reduction Algorithms and Perceptive Preprocessing for Robust Speech Recognition*, in: Psychophysics, Physiology, and Models of Hearing, edited by T. Dau, V. Hohmann, B. Kollmeier, World Scientific, Singapore, pp. 267–270, 1999.

[kle99] M. Kleinschmidt, T. Wittkop, B. Kollmeier: *Evaluation of monaural and binaural speech enhancement for robust auditory–based automatic speech recognition*, J. Acoust. Soc. Am., 105 (2), p. 977. Joint Meeting ASA/EAA/DEGA, Berlin, Germany, 1999.

[nad97] J.-P. Nadal, N. Parga: *Redundancy Reduction and Independent Component Analysis: Conditions on Cumulants and Adaptive Approaches*, Neural Computation, 9, pp. 1421–1456, 1997.

[tch99] J. Tchorz, B. Kollmeier: *A Model of Auditory Perception as Front End for Automatic Speech Recognition*, J. Acoust. Soc. Am., 106 (4), pp. 2040–2050, 1999.