Praktische Aspekte von Mehrmikrofonanordnungen als Frontend für Spracherkenner

Dr. G. Uhlrich; D. Schuchardt; H. Baesekow
ABS Gesellschaft für Automatisierung, Bildverarbeitung und Software mbH, Jena; Email: info@abs-jena.de

Einleitung

Im Rahmen des EU-Forschungsprojektes INSPIRE (**IN**fotainment management with **SP**eech **I**nteraction via **RE**mote-microphones and telephone interfaces) wird gemeinsam mit den Partnern WCL[1] und KNOWLEDGE[2] die Teilaufgabe „Untersuchung von Anordnungen zur akustischen Signalvorverarbeitung in Verbindung mit Spracherkennungssystemen unter realistischen, räumlichen Bedingungen“ bearbeitet. Grundsätzlich bietet sich eine nahezu unüberschaubare Vielfalt von Mikrofonanordnungen für die Zielstellung der Verbesserung des Signalstörabstandes an [1]. Da jedoch als ein Hauptaspekt des oben genannten Forschungsprojektes auch die Kundenakzeptanz zu betrachten ist, muss man sofort erhebliche Einschränkungen unter dem Aspekt der Kosten sowie der Akzeptanz einer räumlich nicht zu voluminösen Mikrofonarrayanordnung treffen. Im Rahmen des Forschungsprojektes sind an mehreren Standorten gleiche Versuchsanordnungen zu errichten, daher spielen selbst die Kosten für den Laboraufbau in allen Überlegungen eine entscheidende Rolle. Deshalb galt es die Frage zu beantworten, wie man als wirtschaftlichen Kompromiss eine Mehrmikrofonanordnung hoher Variabilität realisieren kann. Dabei muss die Flexibilität gegeben sein, verschiedenste algorithmische Ansätze der Sprachsignalverbesserung zu untersuchen. Ein pragmatischer Ansatz zur Herangehensweise und die Schaffung der entsprechenden messtechnischen Voraussetzungen sollen hier als Erfahrungsbericht dargestellt werden. Bewusst werden auch die scheinbar trivialen Probleme erwähnt. Die Autorengruppe betrachtet sich hierbei nicht als Spezialist, sondern als Nachnutzer der bisher wissenschaftlich untersuchten Signalvorverarbeitungsmethoden vom Standpunkt der robusten Spracherkennung aus und möchte ausdrücklich anbieten, neue Ansätze im genannten Anwendungsszenario realitätsnah zu untersuchen.

Aufgabenstellung im Projekt INSPIRE

Das Gesamtziel des INSPIRE-Projektes besteht darin, die Nutzung verfügbarer Spracherkennungstechnologien für einen nutzerfreundlichen Zugang zu Informations- und Unterhaltungsgeräten sowie zur Steuerung sonstiger Hausgeräte im Heimbereich zu ermöglichen und diese Technologien unter dem Gesamtaspekt nutzerfreundlicher Dialoge zu adaptieren. Eine wesentliche Teilaufgabe stellt die Untersuchung von effektiven Möglichkeiten einer Signalvorverarbeitung zur Erhöhung der Robustheit eines Spracherkenners dar.

Zugrunde gelegt wurde hierbei der Einsatz von bis zu 3 Mikrofon-Arrays (zumindest im Labor), welche im Raum irregulär verteilt sind (Abbildung 1).

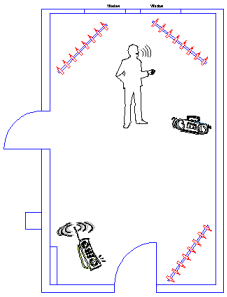

Abbildung 1: Prinzipanordnung für Spracherkennungstests mit 3 Mikrofon-Zeilen-Arrays

[1] Wire Communications Laboratory, Griechenland

[2] Knowledge S. A., Griechenland

Hierbei sind folgende Signalverarbeitungsaufgaben zu implementieren und zu untersuchen:

- automatische Auswahl eines dem Sprecher am nächsten liegenden Mikrofon-Arrays
- Sprecherortung bzw. Sprecherverfolgung (Problem: Wechsel zwischen den Arraybereichen)
- Beamforming:
  - konventionell
  - adaptiv
- Echokompensation (Kompensation des Übersprechens von elektrisch verfügbaren Signalen, welche über Lautsprecher den Raum beschallen wie:
  - Lautsprechersignale des MMI
  - Lautsprechersignale der Audioanlage bzw. des Fernsehers)
- ggf. Störgeräuschunterdrückung (räumlich verteilte Fremdstörer; Störsprecher etc.)
- ggf. Quellentrennung und Enthallung

In [2] findet man eine aktuelle, ernüchternde Einschätzung der momentanen Möglichkeiten bei Echokompensation, Störgeräuschunterdrückung, Quellentrennung und Enthallung unter realistischen Szenarien, welche kurzfristige Erwartungen stark relativieren.

Folgende konkrete Ziele werden im Projekt verfolgt:

- Sammeln von Erfahrungen mit Signalvorverarbeitungsalgorithmen bei Mehrmikrofonsignalen zur Erhöhung der Robustheit eines Spracherkenners.
- Verifikation von aus der Literatur bekannten Ansätzen in einem realen Anwendungsszenario.
- Untersuchung der gegenseitigen Abhängigkeit/Wechselwirkung von Signalvorverarbeitung und Spracherkennungsalgorithmus. Aufgrund bisheriger Erfahrungen gehen wir davon aus, dass es nicht sinnvoll ist, diese Systeme einzeln zu optimieren und dann einfach in Kette zu schalten. Optimale Ergebnisse wird man nur erzielen, wenn man beide Systeme gemeinsam optimiert. So muss in der Regel der Spracherkenner mit Sprachproben trainiert (bzw. adaptiert) werden, welche die vorgesehene Signalverarbeitung durchlaufen haben (sowohl durch Ortungsfehler des Vorverarbeitungs-Algorithmus bedingte Verzerrungen und Artefakte als auch das andere „raumakustische“ Abbild des Sprachsignals).
- Untersuchung der Bewegung des Sprechers im Raum. Hier spielt auch die Robustheit der Sprecherlokalisierung eine wichtige Rolle. Ebenso sollen mit dem vorgestellten System die Fragen der automatischen Umschaltung zwischen zwei Arrays bei Sprecherbewegung untersucht werden.
- Beantwortung der Fragen:
  - Welche Anforderungen müssen an die Mikrofonkapseln wirklich gestellt werden?
  - Welche Toleranzen sind im Vergleich zu den anderen raumakustischen Problemen noch unkritisch?
  - Kann ich ein Mikrofon-Array mit preiswerten, stark tolerierenden Mikrofonen unter normalen raumakustischen Bedingungen einer einfachen Selbstkalibrierung unterziehen?

Auswahl der entsprechenden messtechnischen Voraussetzungen

Raumakustikmessung

Die oben genannten Untersuchungen sollen bei verschiedenen Projektpartnern in realitätsnahen Testräumen erfolgen, welche nicht

zur Umgebung schallisoliert sind. Deshalb ist für die raumakustische Charakterisierung ein MLS-Messsystem unabdingbar.

Es bestand die grundsätzliche Alternative zwischen kompletten Messsystemen von bekannten Anbietern wie Brüel&Kjaer, Norsonic ... und reinen Softwarelösungen für Standard-PC´s. Wegen der Nutzung vorhandener Messmikrofone, Messvorverstärker und der ohnehin zu beschaffenden mehrkanaligen 24-bit-aufgelösten Audiointerfaces wurde die Softwarelösung „Sample Champion PRO 2.5“ eingesetzt [3].

Das Softwarepaket „Sample Champion PRO 2.5“ bietet unter anderem folgende Messmöglichkeiten:

- Impulsantwortmessung mit MLS-Signalen, FFT-Analyse etc.
- MLS-Vorverarbeitung (MLS-Länge, 1 k .... 256 k)
- synchrone Mittelung im Zeit- und Frequenzbereich
- FFT (mit doppelter Genauigkeit) der Impulsantwort oder von „Scope“-Daten
- Echtzeitanalyse (Impulsantwort während Mittelung darstellbar)
- Kalibrierbarkeit des Systems mit interner oder externer Quelle

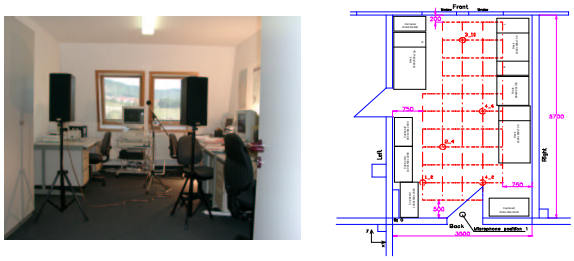

Abbildung 2: Laborraum in Jena mit Messstellenplan

Mikrofon-Array Varianten

Prinzipiell wäre unter der Fokussierung auf die Hauptaufgabenstellung „robuste Spracherkennung“ auch der Einsatz bereits käuflicher Mikrofon-Zeilen-Arrays denkbar und wurde auch ernsthaft mit den Projektpartnern diskutiert. Es sind z. B. folgende Geräte verfügbar*:

Hersteller / Anbieter	Produkt
Acoustic Magic	Voice Tracker ™
Emkay	IntelliSonic ™
Andrea Electronics	DA400

Der Nachteil ist die nicht veränderbare Mikrofonanordnung und die Tatsache, dass die Arrays in der Regel eine eigene Signalverarbeitung enthalten und als „Black-Box“ nur ein Ausgangssignal liefern. Man ist also auf interne Signalverarbeitungsautomatismen angewiesen. Diese sind in der Regel auf den Fall eines dominanten Sprechers im Nahfeld (Sprecherortung und Beamforming) programmiert. Zumindest in Grenzbereichen der sicheren Ortbarkeit wäre eine Kommunikation mit der nachfolgenden Spracherkennungseinheit nötig. Eine automatische Auswahl unter mehreren im Raum verteilten Arrays wäre höchstens für den Fall der jeweiligen Nahbesprechung aufgrund der Pegelverhältnisse möglich. Man muss davon ausgehen, dass für einen optimalen Einsatz von Mikrofon-Arrays als Frontend für Spracherkenner stets ein „intelligentes“ Zusammenwirken und eine wechselseitige Adaption dieser beiden Komponenten nötig sind. Solche Fragestellungen sollen auch unter dem übergeordneten Aspekt eines Nutzerdialoges untersucht werden. Deshalb war die Realisierung eines möglichst variablen Labor-Mikrofon-Arrays mit mehrkanaligem Signal-Interface zum Rechner nötig. Hierzu waren die weiter unten beschriebenen Aspekte zu beachten.

* 2D-Anordnungen sind bisher als handelsübliche Systeme nicht bekannt

Die o. g. handelsüblichen Zeilenarrays sind interessante Vergleichsobjekte. Aufgrund der zukünftig (zu vermutenden) starken Verbreitung entstehen hiermit wohl die geringsten Kosten.

Im Hinblick auf eine spätere Akzeptanz ergeben sich für ein Labor-Mikrofon-Array folgende Grundanforderungen:

- die technische Realisierbarkeit zu akzeptablen Kosten muss absehbar sein
- die prinzipbedingte geometrische Ausdehnung muss in eine Konstruktion überführbar sein, welche zumindest zukünftig akzeptabel erscheint (das Mikrofon-Array muss entweder in ohnehin gegebenen Geräten, Raumeinbauten oder im Baukörper „versteckbar“ sein oder als „Design-Element“ akzeptabel werden).

  → Zeilenanordnungen erscheinen unter diesen Aspekten noch viel eher akzeptabel als Flächenanordnungen oder gar räumliche Anordnungen.
  → Es sollten die Möglichkeiten von Zeilenanordnungen (ggf. auch von Anordnungen in einer rechteckigen Grundfläche) im Raum untersucht werden. Hierbei ist auch der Einsatz von mehreren Zeilenarrays in einem Raum angedacht.

Mehrere Zeilenarrays erlauben prinzipiell:

- die automatische Auswahl des dem Sollsprecher nächsten Arrays (bzw. des am günstigsten fokussierbaren Arrays)
- die Nutzung der nicht auf den Sollsprecher fokussierten Arrays als Referenz für räumlich verteilte Störer (Fokussierung auf Störer). Man beachte hierzu jedoch die kritische Einschätzung der Möglichkeiten einer Geräuschkompensation [1].

Anforderungen an Konstruktion und Fertigung eines Mikrofon-Array

Prinzipiell wäre für die Laboruntersuchung auch der Einsatz einer Arrayanordnung von Messmikrofonen denkbar. Die hierfür entstehenden Kosten stehen nach Ansicht der Projektpartner in keinem Verhältnis zu den zu erwartenden höheren Genauigkeiten und geringeren messtechnischen Problemen. Diese Aussage wird verstärkt, wenn man die begrenzte Signalbandbreite der Sprache und das realitätsnahe Einsatzszenario betrachtet (Preise für eine 8-Mikrofonanordnung inkl. Spannungsversorgung und Vorverstärker von 3.000 bis > 15.000 € sind am Markt üblich). Auch spielt eine entscheidende Rolle, dass bis 24 Mikrofone pro Testraum zum Einsatz kommen und diese Anordnung wiederum an 3 verschiedenen Standorten realisiert wird (d. h. 72 Mikrofone und 72 Signalverarbeitungskanäle sind zu realisieren). Deshalb werden nachfolgend die Anforderungen an ein solches Mikrofonsystem und eine pragmatische, kostengünstige Lösung sowie einige praktische Detailprobleme beschrieben.

Konstruktive Aspekte

- Definition einer geeigneten Mikrofonarrayanordnung (Zeile, Parallel-Array, Quadrat-Array, ...)
- Festlegung der Mikrofonanzahl und Untersuchung von geeigneten Möglichkeiten zur Mikrofonbefestigung / -halterung
- Bestimmung einer optimalen konstruktiven Variante zur flexiblen und reproduzierbaren Positionierung
- Beachtung akustischer Aspekte (Reflexionen, Körperschallentkopplung)
- Positionierung des Vorverstärkers
- Auswahl und Fertigung eines geeigneten Kabelsystems, passend zum Soundkarteninterface
- Beachtung der mechanischen Stabilität des Systems und des Erscheinungsbildes

elektrisch-akustische Anforderungen

- Untersuchungen und Messungen zur Auswahl geeigneter Mikrofonkapseln
- Selektion von möglichst äquivalenten Kapseln (elektrisch und akustisch) aus einer großen Anzahl von Standardkapseln

- Anforderungen an das Spannungsversorgungsmodul definieren
- Gewährleistung einer optimalen Entkopplung von Versorgungs- und Mikrofonsignalen
- Entwicklung eines 8-kanaligen Vorverstärkermoduls
- Festlegung des Interfaces zur Soundkarte (+ 4 dBu, TRS)

Folgende Variante wurde umgesetzt:

Konstruktion

Typ:	2D-Parallel Array, 2 justierbare Stangen als Träger für die Mikrofonhalterung
Anzahl der Mikrofone:	8 beliebig positionierbare Mikrofonkapseln (optional auch mehr Mikrofone möglich)
Positionierbereich:	100 cm in x-Richtung 20 cm in y-Richtung
System:	Array und Mikrofonvorverstärker sind auf einem Mikrofonständer montiert

Abbildung 3: Laboraufbau Mikrofon-Array mit 24k-MotU Sound-Interface

Mikrofone

Kapseln:	entsprechend gemessene und ausgewählte Elektret-Mikrofonkapseln WM54BT (Panasonic) ∅ 9,7 mm
Befestigung:	körperschallentkoppelte Mikrofonhalterung
Konstruktion:	geschirmte Mikrofonkabel mit robusten Spezialsteckern (geringe Abmessungen)

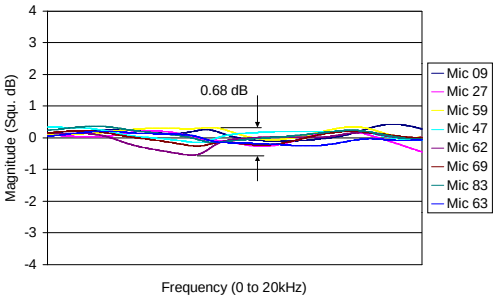

Abbildung 4: Abweichungen der Übertragungsfunktion von 8 ausgewählten Mikrofonkapseln (Frequenzanalyse der Impulsantwort, Direktschallanteil)

Realisierung eines 24-Kanal-Frontend

Prinzipiell stehen auch hier eine Vielzahl von Alternativen zur Verfügung. In die engere Wahl wurden gezogen:

a) 3 x 8 Mikrofon-Array + 3 Digimax + 1 RME Digiface

b) 3 x 8 Mikrofon-Array + 3 Aardvark Q 10 Pro + 3 PCI-Karten

c) 3 x 8 Mikrofon-Array + 3 x 8k-MVV MotU 24 i + PCI-324

Gewählt wurde Variante c, wobei hierfür 24 Vorverstärkerkanäle zu realisieren waren. Es wurde entschieden, eine Eigenentwicklung eines 8-kanaligen Vorverstärkers zu realisieren, welche letztlich zu wesentlich günstigeren Stückkosten, insbesondere auch im Hinblick auf spätere Laborsysteme (für wissenschaftliche Untersuchungen) mit mehr als 24 Kanälen, führt.

Realisierter 8-kanaliger Mikrofon-Vorverstärker

Typ:	8-Kanal-Vorverstärker
Spannungsversorgung:	+/- 12 V
Input:	8 robuste Mehrpolaudiobuchsen (Fabrikat Binder)
Output:	8 Klinkenbuchsen 6,3 mm unbalanced (optional TRS balanced)
Gehäuse:	220 x 140 x 65 mm³

Elektrische Eigenschaften

Übertragungsfrequenzbereich:	20 Hz - 20 kHz (+/- 1dB)
Eingangsempfindlichkeit:	1 mV_{eff} bei SNR > 90 dB
Verzerrung (THD):	< 0,01 % bei U_e = 1 mV_{eff}
Ausgangspegel max.:	2,5 V_{eff} (entspricht + 10 dBu)
Verstärkung:	30 - 60 dB in 4 festen Schritten (einstellbar mittels Jumper)
Verstärkungstoleranz der Kanäle:	< ± 0,5% (entspricht ± 0,04 dB) (bei V = 60 dB)
Eingangsimpedanz:	< 1 kOhm
Ausgangsimpedanz	< 1 kOhm (270 Ohm)
Phantomspeisung:	+ 9 V (schaltbar)
Popfilter (Hochpassfilter):	schaltbar (100 Hz - 3 dB, 30 Hz > -30 dB)

Abbildung 5: 8-kanaliger Mikrofon-Vorverstärker

Zusammenfassung

- Es liegt ein, für die Untersuchung moderner Signalvorverarbeitungs-Ansätze geeignetes, Messequipment mehrfach vor. Entsprechende Einsatzerfahrungen hierzu wurden gesammelt.
- Die Auswahl geeigneter Signalvorverarbeitungs-Varianten ist Gegenstand weiterer Untersuchungen.
- Das Messequipment und die Erfahrungen können nachgenutzt werden, insbesondere liegt ein Ansatz vor, für wissenschaftliche Untersuchungen auch Systeme mit mehr als 24 Kanälen kostengünstig zu realisieren.
- Es wird ausdrücklich darum gebeten, Erkenntnisse anderer Forschungsgruppen zur Verifikation im Zusammenhang mit einer Spracherkennungslösung unter dem geschilderten Einsatzszenario zum Test zur Verfügung zu stellen.

Literatur

- [1] Drews, Martin: Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestörter Sprache, Dissertation, Technische Universität Berlin, 1999
- [2] Kellermann, Walter u.a.: Signalverarbeitung für akustische Mensch/Maschine-Schnittstellen. In: „Elektronische Sprachsignalverarbeitung“, Band 24, Rüdiger Hoffmann. Dresden: w.e.b. Universitätsverlag, 2002
- [3] Guidorzi, Paolo: Software Sample Champion PRO 2.5 User Guide, PureBits.com, 2000

Diese Seite ist mit Absicht unbedruckt.

This page intentionally left blank.