Projekt Amigo - Sprachsignalverarbeitung im vernetzten Haus

Jörg Schmalenströer[1], Reinhold Häb-Umbach[2], Ernst Warsitz[3]

Fachgebiet Nachrichtentechnik Universität Paderborn, 33098 Paderborn, Deutschland

[1] *schmalen@nt.uni-paderborn.de* [2] *haeb@nt.uni-paderborn.de* [3] *warsitz@nt.uni-paderborn.de*

Einleitung

Ziel des integrierten Projektes Amigo ("Ambient Intelligence for the networked home environment") im 6. EU Rahmenprogramm ist die Verbindung der Bereiche Unterhaltungselektronik, Haushaltsgeräte, mobile Kommunikationsgeräte und Computer in der vernetzten Hausumgebung. Hierzu wird eine quelloffene, standardisierte und interoperable Middleware zur Verbindung von Geräten und Bereitstellung von Diensten entwickelt [1].

Amigo ist als Kontext bewusstes System mit einer intelligenten, multi-modalen Benutzerschnittstelle angelegt.

Neben expliziten Benutzereingaben, wie Sprach- und Gestenerkennung, wird das System auch implizite Informationen gewinnen und auswerten. Metadaten über Themengebiete bei Gesprächen, Nutzergewohnheiten und soziale Komponenten können somit für neue Dienste verwendet werden.

Sprache als natürlichste Art der Mensch-zu-Mensch Kommunikation, bildet eine der Kernmodalitäten der Amigo Benutzerschnittstelle. Unsere Aufgabe hierbei ist neben der Aufnahme und Auswahl der akustischen Signale, der Störunterdrückung und Echounterdrückung auch die Bestimmung von impliziten Benutzereingaben aus den akustischen Signalen. Diese Kontextinformationen werden in der akustischen Szenenanalyse aus den Audiosignalen gewonnen. Die bekannte Fragestellung "Wer spricht Wann" wird hierbei von uns um die Positionsinformation "Wo" erweitert. Metadaten sind hierbei Sprachaktivitätsdetektion, Sprecherwechselerkennung, Sprecherpositionsschätzung und Sprecheridentifikation.

Randbedingung für Algorithmen im Amigo System ist eine näherungsweise echtzeitfähige Laufzeit der Algorithmen bei niedrigen Latenzzeiten. Die von uns eingesetzten Verfahren kombinieren Informationen der verschiedenen Komponenten der akustischen Szenenanalyse zur Verbesserung der Metadaten.

Akustische Positionsschätzung

Die akustische Positionsschätzung basiert auf Mikrophongruppen, die aus linear angeordneten Fernfeld Mikrophonen im Abstand von jeweils 5cm bestehen. Für eine optimale Positionsschätzung sollten die Mikrophongruppen gleichmäßig im Raum verteilt werden, wobei der Einfluss von glatten, stark reflektierenden Flächen berücksichtigt werden muss. Die Signale einer jeden Mikrophongruppe werden in einem *Filter-and-Sum-Beamformer* (FSB) einer adaptiven Strahlformung unterzogen [2]. Nebenprodukt der Strahlformung ist die Möglichkeit, aus den Filterkoeffizienten des FSB eine Richtungsschätzung (Direction-of-Arrival, DoA) vornehmen zu können [3]. Somit kann für jede Mikrophongruppe $\texttt{FSB}_i$ eine DoA-Schätzung $\widehat{\alpha}_i$, mit einem dazugehörigen Konfidenzwert κ_i, gewonnen werden. Während der Positionsschätzung werden aus den DoA Informationen der Mikrophongruppen und der geometrischen Anordnung der Gruppen im Raum Schnittpunkte (x_{ij}, y_{ij}) berechnet. Die Positionsschätzung ergibt sich anschließend aus dem mit den Konfidenzwerten gewichteten Schwerpunkt der Schnittpunkte (siehe Abb. 1).

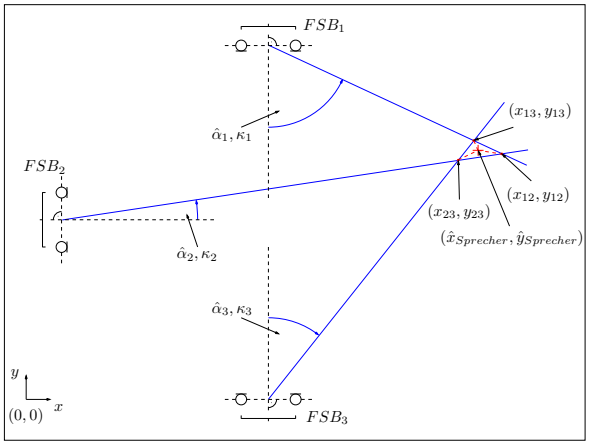

Abbildung 1: Positionsbestimmung mit 3 Mikrophongruppen

Sprecherwechselerkennung und Sprecheridentifikation

Die Sprecheridentifikation setzt zwingend eine robuste Sprecherwechselerkennung voraus. Hierzu können verschiedene Ansätze wie z.B. BIC (Bayesian Information Criterion) genutzt werden [4]. Grundsätzlich bleibt aber das Problem der schlechten Signal-zu-Rausch-Verhältnisse der entfernten Mikrophone und die zusätzlich störenden Echos. Anders als in der klassischen Annotation von Sprachdaten (wie z.B. bei Fernsehaufzeichnungen) steht jedoch durch den Einsatz von Mikrophongruppen die zusätzliche Information der Sprecherposition zur Verfügung. Unter der üblicherweise gegebenen Annahme, dass ein aktiver Sprecher seine Position nur langsam ändert und verschiedene Sprecher genügend räumlich getrennt sind, kann eine robustere Sprecherwechselerkennung durchgeführt werden [5].

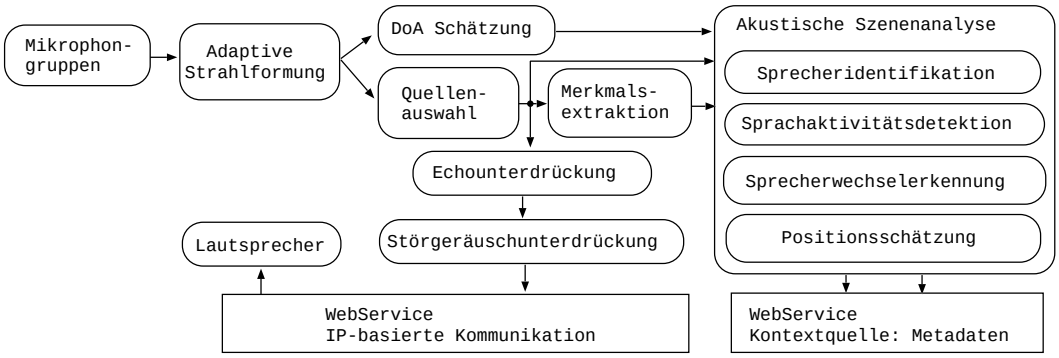

Abbildung 2: Akustische Szenenanalyse

Akustische Szenenanalyse

Das System zur akustischen Szenenanalyse ist in Abb. 2 dargestellt. Zunächst werden die mehrkanaligen Audiosignale durch die adaptive Strahlformung zu einem einkanaligen Signal zusammengefasst. Anschließend findet die DoA Schätzung für die Sprecherlokalisation statt. Falls mehrere Mikrophone bzw. Mikrophongruppen vorhanden sind, wird das qualitativ beste Signal der zur Verfügung stehenden Audiosignale in der Quellenauswahl bestimmt und an die Merkmalsextraktion weitergegeben. Die Merkmalsextraktion führt zunächst eine 2-stufige Wiener Filterung mit anschließender Berechnung der MFCC-Merkmale durch. Zusätzlich wird für die Sprecheridentifikation ein Merkmal für die Stimmhaftigkeit berechnet.

Die DoA Schätzwerte und die Merkmalsvektoren werden dann in der akustischen Szenenanalyse durch einen Viterbi Algorithmus gemeinsam verarbeitet. Die hierbei gewonnenen Metadaten werden als Kontextquelle im Amigo System bereitgestellt.

Kontextquellen

Im Amigo System werden Kontextinformationen durch Kontextquellen, mit standardisierter WebService Schnittstelle, bereitgestellt. Die Schnittstelle bietet hierzu den synchronen Zugang durch die Methode "query" und den asynchronen Zugang durch die Methoden "subscribe/unsubscribe" an. Alle Anfragen werden in SPARQL (SPARQL Protocol and RDF Query Language) formuliert und mit einer RDF (Resource Description Framework) Beschreibung der Metadaten beantwortet. Durch die konsequente Verwendung von WebServices und XML kodierten Daten und der Einführung von Ontologien zur Kontextbeschreibung wird eine hohe System-Interoperabilität erzielt.

Zeitliche Anforderungen

Das Amigo System hat das Ziel, den Benutzer in nahezu Echtzeit zu unterstützen. Dies hat zur Folge, dass alle verwendeten Algorithmen in der Sprachsignalverarbeitung und in der akustischen Szenenanalyse zwei Kriterien erfüllen müssen. Zum einen müssen sie mit beschränkten Ressourcen auskommen, da eine Vielzahl von Komponenten sich eine begrenzte Anzahl an CPUs teilen muss. Zum anderen müssen die Ergebnisse der akustischen Szenenanalyse nach einer möglichst kurzen Zeit vorliegen um dem Echtzeitanspruch des Systems gerecht zu werden. Weiterhin ist es nötig, dass die ausgewählten Algorithmen auf fortlaufend eintreffende Audiodatenströme angewandt werden können und nicht die Voraussetzung haben, dass der zu bearbeitende Datensatz vollständig ist.

Ein Grundidee der Sprachsignalverarbeitung im Amigo System ist die konsequente Aufteilung der Audioverarbeitung in parallele Threads, welche durch Interprozesskommunikation (IPC) verbunden sind. Hierdurch wird eine Trennung zwischen verzögerungskritischen Prozessen, wie z.B. Audiokommunikation, und Prozessen mit niedrigeren Anforderungen an die Latenz, wie z.B. Sprecheridentifikation, geschaffen.

Danksagung

Diese Arbeit wurde unterstützt durch das Projekt der europäischen Union "Amigo - Ambient intelligence for the networked home environment" [1].

Literatur

[1] Amigo - Ambient Intelligence for the networked home environment, "http://www.amigo-project.org", 2007

[2] E. Warsitz, R. Haeb-Umbach, "Acoustic Filter-and-Sum Beamforming by Adaptive Principal Component Analysis", Proc. ICASSP05, Philadelphia, USA, March 2005

[3] E. Warsitz, R. Haeb-Umbach, S. Peschke, "Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization", Proc. ICSLP 2004, Jeju, Korea, 2004

[4] S. Tranter, D. Reynolds, "An Overview of Automatic Speaker Diarization Systems", IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 5, Sep. 2006

[5] J. Schmalenstroeer, R. Haeb-Umbach, "Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming", Proc. Interspeech 2006, Pittsburgh, USA, 2006