Multichannel Dereverberation and Noise Reduction for Hands-Free Speech Communication Systems

Dominic Schmid

Moderne Kommunikationssysteme wie Mobiltelefone, Telekonferenzeinrichtungen oder Sprachdialogsysteme dienen der Übertragung und Verarbeitung von Sprachsignalen. Um den Bedienkomfort zu steigern, sind diese Systeme häufig mit einer Freisprechfunktion ausgestattet, die es dem Benutzer erlaubt, sich in einem bestimmten Abstand von den Mikrofonen frei zu bewegen. In den meisten Aufnahmesituationen werden neben dem gewünschten Sprachsignal auch Hintergrundgeräusche und Hall, d.h. Reflexionen des ursprünglichen Signals an Wänden, Decken oder naheliegenden Objekten, aufgezeichnet. Beide Störungen verschlechtern nicht nur die Sprachqualität und die Sprachverständlichkeit, sondern auch die Erkennungsraten von automatischen Spracherkennungssystemen. Ziel dieser Arbeit war daher die Entwicklung mehrkanaliger Sprachverbesserungsalgorithmen, die das ungestörte Sprachsignal aus den gestörten Mikrofonsignalen rekonstruieren.

Der erste Teil der Arbeit befasst sich zunächst mit den Voraussetzungen, unter denen eine Schätzung des Quellensignals aus den gestörten Mikrofonsignalen möglich ist. Durch die Gegenüberstellung bekannter Verfahren aus der Literatur werden deren Gemeinsamkeiten aufgezeigt und die Filterlängen untersucht, für die eine geschlossene Lösung des jeweiligen Schätzproblems existiert. Eine genauere Analyse des Least-Norm-Entzerrers zeigt, dass das Quellensignal durch eine zweistufige Filterstruktur rekonstruiert werden kann, die aus einem Matched-Filter-Array und einem einkanaligen inversen Filter besteht. Da die Schätzung des Quellensignals die Kenntnis der Raumimpulsantworten zwischen dem Sprecher und den jeweiligen Mikrofonen voraussetzt, werden anschließend bestehende Verfahren zur blinden Kanalidentifikation analysiert. Dabei wird nachgewiesen, dass diese Algorithmen in der Gegenwart von naheliegenden Kanalnullstellen und Beobachtungsrauschen zu systematischen Schätzfehlern führen, die sich als einkanalige Filterfehler modellieren lassen. Um auch unter schwierigen akustischen Bedingungen eine Evaluation dieser Verfahren zu ermöglichen, wird ein neues Kanalabstandsmaß vorgeschlagen, das diese gemeinsamen Filterfehler berücksichtigt.

Im zweiten Teil der Arbeit werden neue adaptive Sprachverbesserungsalgorithmen vorgestellt, die sowohl Hall als auch Hintergrundgeräusche reduzieren. Die Verfahren verwenden ein rahmenbasiertes Beobachtungsmodell im Frequenzbereich, das die Übertragung des Quellensignals durch die hallige und gestörte akustische Umgebung beschreibt. Der unbekannte Kanalvektor wird dabei als eine Zufallsgröße modelliert, die einem Markov-Modell erster Ordnung folgt. Zusammen mit den Beobachtungsgleichungen ergibt sich so die Zustandsraumdarstellung eines dynamischen Systems, das explizit die Zeitvarianz der Übertragungskanäle berücksichtigt. Ausgehend von diesem Modell werden drei iterative Verfahren vorgestellt, die mit Hilfe des Expectation-Maximization-Algorithmus die unbekannten Modellgrößen schätzen. Die Ansätze unterscheiden sich in der Modellierung des Quellensignalvektors, der als deterministischer Parameter, als Parameter mit Prior-Verteilung und als Zufallsgröße beschrieben wird, um so Maximum-Likelihood, Maximum-A-Posteriori und Variational-Bayes'sche Schätzverfahren herzuleiten. Der Expectation-Schritt dient dabei der Schätzung der Posterior-Verteilung der gewählten Zufallsvariablen, während der Maximization-Schritt Punktschätzungen der Modellparameter liefert. Durch die sequentielle Ausführung beider Schritte ergeben sich adaptive Sprachverbesserungsalgorithmen, die eine rekursive Kanalschätzung und eine mehrkanalige Entzerrung kombinieren.

Eine umfangreiche Evaluation der neuen Verfahren mit Hilfe instrumenteller Maße verdeutlicht, dass diese die Sprachqualität für eine Vielzahl realistischer Hall- und Geräuschszenarien verbessern und sich auch auf zeitveränderliche akustische Umgebungen einstellen können. Versuche mit einem automatischen Spracherkennungssystem zeigen darüber hinaus, dass eine Verarbeitung der Mikrofonsignale mit den vorgestellten Algorithmen die Erkennungsraten deutlich steigert. Zum Abschluss der Arbeit wird die Implementierung auf einer Echtzeitplattform beschrieben. Das System lässt sich über eine grafische Benutzeroberfläche bedienen und erlaubt dem Benutzer eine Bewertung der prozessierten Signale in Echtzeit. Die durchgeführten Experimente bestätigen, dass die entwickelten Verfahren auch unter realen akustischen Bedingungen eine Verbesserung der Sprachqualität erreichen.