Vergleich und Optimierung von akustischen Merkmalsystemen für die automatische Erkennung von Umgebungssituationen

Ulrich Kordon, Thomas Hutschenreuther

Technische Universität Dresden, Institut für Akustik und Sprachkommunikation, 01062 Dresden, Deutschland, Email:ulrich.kordon@ias.et.tu-dresden.de

1. Einleitung

Verfahren zur akustischen Context-Erkennung haben in den vergangenen Jahren zunehmend an Bedeutung gewonnen. Das wissenschaftliche Interesse war dabei bisher vor allem auf die Auswahl und Optimierung geeigneter Klassifikationsansätze gerichtet. Für die dabei benötigten Objektbeschreibungen wird größtenteils auf Merkmalsysteme zurückgegriffen, die sich in ähnlichen Erkennungsaufgaben zwar als ausreichend tragfähig erwiesen haben (z. B. [1]), wobei aber eine Einschätzung der Leistungsfähigkeit alternativer Verfahren unter vergleichbaren Bedingungen auf Grund der unterschiedlichen Randbedingungen kaum möglich ist. Der Beitrag befasst sich deshalb mit einem Vergleich verschiedener akustischer Merkmalsysteme und der Optimierung der entsprechenden Analysebedingungen unter ansonsten äquivalenten Bedingungen.

2. Datenbasis

Als Datenbasis wurden Aufnahmen von folgenden typischen Geräuschsituationen aus den Bereichen Gesellschaftsräume bzw. Verkehrsmittel verwendet:

Datensatz 1 (Dauer jeweils 10 Minuten):
Mensa (Speisesaal), Hörsaal (Vorlesung), Studentenclub („Party“), PKW-Innenraum (in Fahrt), Straßenbahn-Innenraum (in Fahrt), Verkehrslärm (verkehrsreiche Straße)
Datensatz 2 (Dauer jeweils 20 Minuten):
A) Mensa - Speisesaal, B) Mensa – Außenbereich, C) Mensa – Cafeteria, D) Mensa – Eingangsbereich, E) Mensa – Essenausgabe

Der Datensatz 1 umfasst verschiedenartige Geräusche, um die Merkmalsysteme optimieren zu können. Datensatz 2 weist dagegen sehr ähnliche Geräusche auf, auf deren Basis die Diskriminationsfähigkeit des Klassifikationssystems mit den optimierten Merkmalen bewertet werden soll.

Die Signale wurden mit einer Abtastfrequenz von 32 kHz und einer Auflösung von 16 Bit digitalisiert und in Abschnitte, deren Länge von jeweils 15 – 60 s mit einer Schrittweite von 15 s variiert wurde, unterteilt. Jeder dieser Abschnitte stellt eine Realisierung der entsprechenden Geräuschklasse dar. Damit standen je Klasse zwischen 10 und 40 bzw. 20 und 80 Realisierungen (Datenbasis 1 / Datenbasis 2) zur Verfügung.

3. Vorverarbeitung und Merkmalableitung

Vorverarbeitung und Primäranalyse: Jede dieser Realisierungen wurde zunächst mit einer primären Merkmalvektorfolge beschrieben, die durch eine fortlaufende Kurzzeitanalyse der jeweiligen Realisierung ermittelt wurde. Die Dauer der mit Hamming-Fenster gewichteten Zeitfenster betrug dabei 1 s.

Als primäre Merkmalsysteme kamen folgende Varianten zur Anwendung:

System 1: Mel-Frequency-Cepstral-Coefficients (MFCC):
Nach einer DFT und mel-Skalierung auf Basis einer 67-kanaligen mel-Filterbank wurden durch Logarithmierung und diskreter Cosinus-Transformation der Filterspektren 13 MFCC ermittelt.
System 2: Linear-Prediction-basierte Merkmale (LPC):
Aus dem mit einer Eckfrequenz von 500 Hz Tiefpass-gefilterten Signal wurden zunächst 32 LPC-Koeffizienten extrahiert. Nach Auffüllung mit 0 auf 160 Werte lieferte eine DFT dieses erweiterten Koeffizientensatzes und anschließender gewichteter Inversion der Spektralkoeffizienten den 160-dimensionalen Merkmalsatz.
System 3: Hybrider Merkmalsatz (MIX):
Der hybride Merkmalsatz umfasste die 8 Komponenten spektraler Median, Effektivwertquadrat, zweites Maximum der Autokorrelationsfunktion sowie 5 spektrale Formmerkmale (Frequenz des absoluten spektralen Maximums, Frequenzabstand zwischen den ersten beiden lokalen spektralen Maxima, Anstieg zwischen absoluten und ersten bzw. zweiten lokalen spektralen Maximum sowie Quotient zwischen spektralen Maximum und spektralen Mittelwert).

Während mit den Systemen 1 und 2 Merkmale ausgewählt wurden, die sich bei ähnlichen Problemstellungen wie z. B. der Sprachsignalverarbeitung als entsprechend leistungsfähig herausgestellt haben, sind im System 3 Merkmale zusammengefasst, die nach Auswertung von Signaldarstellungen der verwendeten Datenbasis durch einen humanen Experten eher heuristisch gewonnen wurden.

Sekundäranalyse: Diese für alle Realisierungen ermittelten primären Beschreibungen wurden durch Deltamerkmale in Form der ersten Ableitung der jeweiligen Merkmalvektorfolge ergänzt. Eine Hauptkomponentenanalyse lieferte schließlich die sekundären Merkmalvektorfolgen unterschiedlicher Dimension (MFCC: 5–21 / Schrittweite 2, LPC: 5 / 7 / 9, MIX: 5–19 / Schrittweite 2), um auch den Einfluss der Merkmalvektor-Dimension untersuchen zu können.

4. Klassifikation

Die Erkennung erfolgte mit Hilfe eines Klassifikationssystems auf der Basis stochastischer Markov-Graphen (SMG) (Abbildung 1), das für die Verarbeitung sprachlicher und nichtsprachlicher akustischer Objekte entwickelt wurde [2]. Insgesamt standen mit den Datenbasen 1 und 2, abhängig von der Realisierungslänge (vgl. 2.), 60–240 bzw. 100– 400 Muster zur Verfügung.

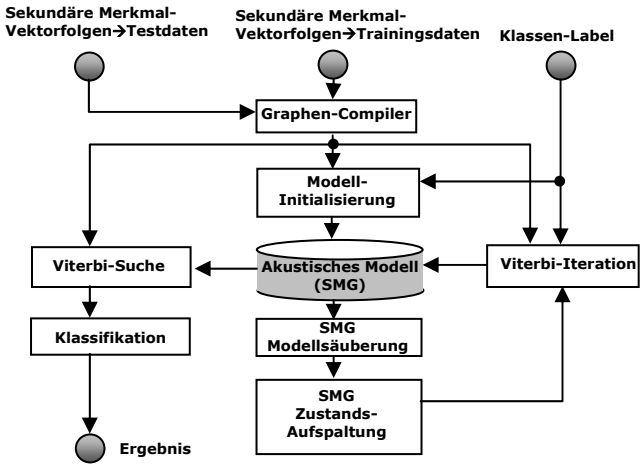

Abbildung 1: Struktur des Klassifikationssystems

Davon wurden jeweils 80 % für das Training und 20 % für die Erkennung verwendet, so dass nach entsprechendem Umlauf *alle* Elemente der Datenbasen als Trainingsdaten *und* Testdaten eingingen.

5. Ergebnisse

Datenbasis 1 - Dimension sekundäre Merkmalvektorfolge: In Abbildung 2 sind die erreichten Erkennungsergebnisse in Abhängigkeit von der Dimension der sekundären Merkmalvektorfolge für eine Realisierungslänge von 30 s dargestellt.

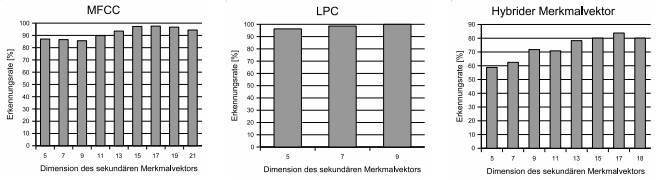

Abbildung 2: Abhängigkeit von der Dimension der sekundären Merkmalvektorfolge

Das Absinken der Erkennungsrate bei höheren Dimensionen kann auf den dann geringeren Umfang der Trainingsdaten zurückgeführt werden. Als Kompromiss zwischen Erkennungsrate und Aufwand ergaben sich Werte von 9 … 12.

Datenbasis 1 - Realisierungslänge: Die Abhängigkeit der Erkennungsrate von der Länge der Realisierung zeigt Abbildung 3.

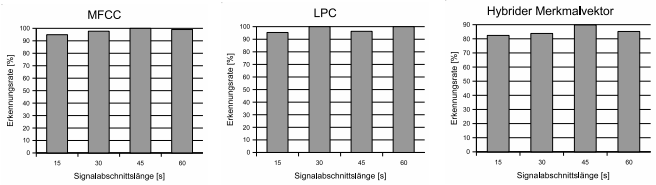

Abbildung 3: Abhängigkeit von der Realisierungslänge

Die für die Dimensionalität wurde der jeweils verfahrensspezifische optimale Wert (MFCC – 9; LPC, MIX – 17, vgl. Abb. 2) gewählt. Auch hier muss beachtet werden, dass die Trainingsdatenmenge mit steigender Realisierungslänge abnimmt, was ggf. das Absinken der Erkennungsrate bei höheren Werten erklärt. Als optimal kann ein Wert von 30 s angenommen werden.

Datenbasis 1 - Modellkomplexität: Der Einfluss der Modellkomplexität (Zustandszahl des HMM = $2^{\text{Anzahl Aufspaltungen}}$, vgl. [2]) ist in Abbildung 4 angegeben.

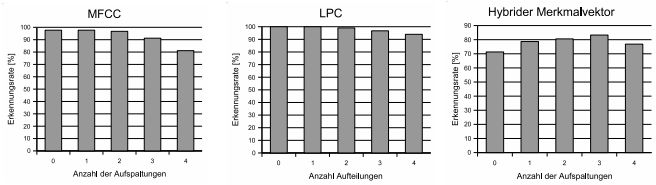

Abbildung 4: Abhängigkeit von der Modellkomplexität

Für MFCC und LPC können 2…4, für MIX 4…8 Zustände als Optimum angesehen werden.

Unter Verwendung der jeweils optimalen Werte wurden mit der Datenbasis 1 für LPC 100 %, MFCC 97 % und MIX 84 % Erkennungsrate erreicht.

Datenbasis 2: Für die akustisch sehr ähnlichen Realisierungen der Datenbasis 2 konnten unter Optimalbedingungen (LPC, Dimension 9, Realisierungslänge 30 s, 2 Zustände) die in Tabelle 1 aufgeführten Erkennungsergebnisse erreicht werden.

Tabelle 1: Erkennungsergebnisse für Datenbasis 2

Klasse	A	B	C	D	E
A	**1,00**	**0,15**			**0,08**
B		**0,85**			
C			**0,89**	**0,10**	
D			**0,11**	**0,90**	
E					**0,92**

Als mittlere Gesamterkennungsrate ergab sich damit ein Wert von 91,1 %.

6. Schlussfolgerungen

Die für vergleichbare akustische Mustererkennungsaufgaben geeigneten Merkmale weisen auch für die akustische Context-Erkennung optimale Eigenschaften auf. Bereits bei relativ geringer Dimension der Merkmalvektoren von ca. 10 und Realisierungslängen von ca. 30 s werden stabile Erkennungsergebnisse auch bei akustisch sehr ähnlichen Mustern erreicht [3].

Weitere Untersuchungen sollten auf die Steigerung der Robustheit der Erkennung bei komplexeren und umfangreicheren Datenbasen gerichtet sein. Möglichkeiten bestehen dazu in der weiteren Optimierung des hybriden Merkmalsystems und der HMM-Struktur.

Literatur

[1] Peltonen, V.: Computational Auditory Scene Recognition. Diss. Tampere University of Technology, 2001

[2] Eichner, M.: Spracherkennung und Sprachsynthese mit gemeinsamen Datenbasen. Diss. TU Dresden – IAS, 2008.

[3] Hutschenreuther, T.: Automatische Erkennung von Umgebungssituationen anhand von akustischen Merkmalen. Studienarbeit, TU Dresden – IAS, 2007.