Universität für Musik und darstellende Kunst Graz

Technische Universität Graz

Fraunhofer Institut für integrierte Schaltungen Erlangen

Diplomarbeit

Optimierung eines Systems zur automatischen Mehrkanaltonerweiterung von TV- und Filmton

Vorgelegt von:	Patrick Gampp
Matrikelnummer:	9931027
Abgabe:	Januar 2011
Studiengang:	Elektrotechnik Toningenieur
Institut:	Institut für Elektronische Musik und Akustik
Betreuer der Universität:	O. Univ.- Prof. Dr. Robert Höldrich
Betreuer von Fraunhofer:	Dr.- Ing. Christian Uhle
	M.A. Falko Ridderbusch
	Dipl.- Ing. Oliver Hellmuth

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am ……. ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

……. ……..
date (signature)

Kurzfassung

Mit der hohen Marktakzeptanz der Digital Versatile Disc (DVD) seit ihrer Einführung im Jahre 1995, kam es zu einer großen Verbreitung von mehrkanalfähigen Wiedergabesystemen in Privathaushalten.

Ein beträchtlicher Teil der heute erhältlichen Medien im Bereich Musik wird jedoch nicht in Mehrkanaltonformaten produziert. Auch TV-Sendeinhalte wie Serien und ältere Filme sind zum Großteil nur in Zweikanalstereo verfügbar.

Um Medien, die in Zweikanalstereo vorliegen, auch in Verbindung mit mehrkanalfähigen Wiedergabesystemen nutzen zu können, wurde am Fraunhofer IIS ein System zur automatischen Mehrkanaltonerweiterung (Englisch: Upmix) für Musik entwickelt.

Im Rahmen dieser Arbeit wurde dieses System im Hinblick auf die Wiedergabe von TV- und Filmton angepasst. Ein Gestaltungskriterium von großer Bedeutung ist hierbei die klanglich unverfälschte Wiedergabe von Sprache aus dem Centerkanal.

Der hier vorgeschlagene Ansatz sieht vor, Klangparameter des Upmixers über die Zeit zu verändern. Mit Hilfe einer Sprachdetektion soll bestimmt werden, zu welcher Zeit das Eingangssignal des Upmixers Sprache enthält. Auf Grundlage der ermittelten Sprachsegmentgrenzen soll daraufhin ein Übergang zwischen zwei Klangeinstellungen des Upmixers stattfinden, die für die Wiedergabe von Sprache bzw. Musik, Atmosphären usw. angepasst sind.

Zunächst wurde ein System zur Mustererkennung speziell für die Detektion von Sprache in TV- und Filmton angepasst. Die Erweiterungen beinhalten eine Vorverarbeitung der Signale mit Hilfe einer spektralen Gewichtung. Darüber hinaus wurden Stereomerkmale definiert, die Interkanalkohärenz- sowie Interkanalpegeldifferenz-Eigenschaften

des Signals beschreiben. Es wurde eine Nachverarbeitung entwickelt, die einen zusätzlichen Klassifizierer zur Laufzeit trainiert. Schließlich wurde eine Hüllkurvensegmentierung mittels adaptiver Hintergrundpegelberechnung zur Nachverarbeitung der geschätzten Sprachsegmente implementiert.

Es wurden verschiedene Algorithmen zur Berechnung der Steuerfunktion der Klangparameter des Upmixers implementiert und getestet.

Es wurde durch Hörtests gezeigt, dass die Wiedergabequalität von Sprache durch die im Rahmen der Arbeit entwickelten Erweiterungen signifikant verbessert werden konnte. Im Vergleich zu den verwendeten statischen Klangeinstellungen des Upmixers, konnte durch die Überblendung zwischen Klangeinstellungen eine signifikante Verbesserung hinsichtlich der Wiedergabequalität von Sprache bzw. der Breitendarstellung erreicht werden.

Die Überblendungen ausgewählter, kritischer Testsignale waren für mehrere erfahrene Hörer nicht wahrnehmbar. Daraufhin wurde dieser Hörtest mit ausgebildeten Tonmeistern durchgeführt. Diese nahmen die Überblendungen in den meisten Fällen überhaupt nicht, oder als nicht störend wahr.

Abstract

With consumer acceptance of the Digital Versatile Disc (DVD), launched in 1995, surround sound systems have been widespread in private households.

A majority of today's music however, is not produced in multichannel format. TV content such as television series and old films are only available in two-channel stereo audio.

In order to utilize media with two-channel audio with surround sound systems, a blind-upmix-system focusing on playback of music content was developed at Fraunhofer IIS.

This thesis discusses how the upmix-system was adapted for playback of TV and movie-content. An important design criterion was to achieve pristine sound playback of speech coming from the center channel.

The basic concept of the proposed approach consists of varying the sound parameters of the upmixer over time. A speech detection system determines at which time the input signal of the upmixer contains speech. On the basis of these speech segments, a fade between two sound settings is executed. The settings are tuned to the playback of speech and music, atmospheres respectively.

First, a pattern recognition system was adopted especially for the detection of speech in TV- and movie-audio. The developed additions comprise a pre-processing of the signals with the help of spectral weighting. Additionally, stereo features were defined to utilize interchannel coherence and interchannel level differences of the signal. Post-processing was designed, to use an additional classifier which is trained at runtime. Finally, envelope segmentation with adaptive background level calculation for the post-processing of

estimated speech segments was designed and implemented.

Several algorithms for the computation of a control-function of the upmixer's sound parameters were implemented and tested.

Listening tests showed that the quality of speech playback was significantly improved by the developed additions. Furthermore, it was shown, that sound performance with respect to the positioning of sound sources and sound quality of speech can be improved significantly by fading between two sound settings instead of remaining in a single static sound setting.

The fading between two sound settings was not perceived by several experienced listeners. Listening tests were also carried out by experienced sound-engineers, who either did not perceive the fading at all, or perceived it to be of minimal annoyance in most cases.

Danksagung

Einen großen Dank möchte ich Christian Uhle, Falko Ridderbusch und Oliver Hellmuth für ihre Betreuung während dieser Arbeit aussprechen.

Ebenso möchte ich Prof. Robert Höldrich für seine Betreuung seitens der Universität in Graz danken.

Ich möchte meinen Kollegen am Fraunhofer IIS danken. Ich danke meinen Freunden, besonders Christian Göttlinger, Maximilian Huber und Mareike Wieth sowie meiner Familie.

Ich danke meiner Freundin Xiao für ihre großartige Unterstützung.

Einen ganz besonderen Dank möchte ich meinen Eltern aussprechen, die mich meinen Weg gehen ließen und mich dabei in jeder Hinsicht immer unterstützt haben. Ich danke Euch!

Inhaltsverzeichnis

1 Einleitung

Mit der hohen Marktakzeptanz der Digital Versatile Disc (DVD) seit ihrer Einführung im Jahre 1995 [1], kam es zu einer großen Verbreitung von mehrkanalfähigen Wiedergabesystemen in Privathaushalten.

Ein beträchtlicher Teil der heute erhältlichen Medien im Bereich Musik wird jedoch nicht in Mehrkanaltonformaten produziert. Auch TV-Sendeinhalte wie Serien und ältere Filme sind zum Großteil nur in Zweikanalstereo verfügbar.

Um Medien, die in Zweikanalstereo vorliegen, auch in Verbindung mit mehrkanalfähigen Wiedergabesystemen nutzen zu können, wurde am Fraunhofer IIS ein System zur automatischen Mehrkanaltonerweiterung (Englisch: Upmix) für Musik entwickelt [2].

Im Rahmen dieser Arbeit soll das System im Hinblick auf die Wiedergabe von TV- und Filmton angepasst werden. Ein Gestaltungskriterium von großer Bedeutung ist hierbei die klanglich unverfälschte Wiedergabe von Sprache aus dem Centerkanal [3]. Ein Ziel dieser Arbeit ist Entwicklung von Algorithmen, die eine Verbesserung der Wiedergabe bezüglich dieses Kriteriums ermöglichen.

Der hier vorgeschlagene Ansatz sieht vor, Klangparameter des Upmixers über die Zeit zu verändern. Mit Hilfe einer Sprachdetektion soll bestimmt werden, zu welcher Zeit das Eingangssignal des Upmixers Sprache enthält. Auf Grundlage der ermittelten Sprachsegmentgrenzen soll daraufhin ein Übergang zwischen zwei Klangeinstellungen des Upmixers stattfinden. Eine Klangeinstellung ist speziell für die Wiedergabe von Sprache, eine weitere Klangeinstellung für die Wiedergabe von Musik, Atmosphären, Geräuschen usw. angepasst.

Die Arbeit ist folgendermaßen aufgebaut: In Kapitel 2 werden die für das Verständnis

der Arbeit erforderlichen Grundlagen behandelt. Kapitel 3 beschreibt Aufbau und Funktionsweise der Sprachdetektion und Upmixverfahren. Die entwickelten algorithmischen Erweiterungen werden in Kapitel 4 vorgestellt. In Kapitel 5 werden die Ergebnisse der entwickelten Erweiterungen vorgestellt und diskutiert.

2 Grundlagen

2.1 Mustererkennung

Mustererkennung ist der Vorgang durch Sensoren aufgenommene Muster aufgabenspezifisch einer Kategorie zuzuordnen. Der Mensch schafft es scheinbar mühelos Sinneswahrnehmungen in Form von Kategorien zu verarbeiten. Diese Fähigkeit ermöglicht u.a. die Erkennung von Gesichtern, das Verständnis von Sprache oder das Verständnis von Texten. Dieses Kapitel soll einen Überblick über Aufbau (siehe Abb. 2.1) und Funktionsweise der von Maschinen durchgeführten statistischen Mustererkennung geben [4, 5].

2.1.1 Trainingsphase

Viele Mustererkennungsprobleme sind so komplex, dass das System im Vorfeld mit Hilfe von Trainingsdaten auf die Anwendung eingestellt wird. Die inneren Parameter des Mustererkennungssystems werden so variiert, dass der Klassifikationsfehler während der darauf folgenden Arbeitsphase möglichst klein wird. Durch das Training findet also ein Lernprozess statt. Es kann grundsätzlich zwischen drei Lernstrategien unterschieden werden:

- Überwachtes Lernen: Die Klassenzugehörigkeit eines jeden Musters ist während der Trainingsphase bekannt.
- Unüberwachtes Lernen: Die Klassenzugehörigkeit der Muster muss über die Bildung von Gruppen (Englisch: Clustering) im Merkmalsraum, abhängig von den innerhalb des Clustering-Algorithmus gemachten Vorgaben, ermittelt werden.

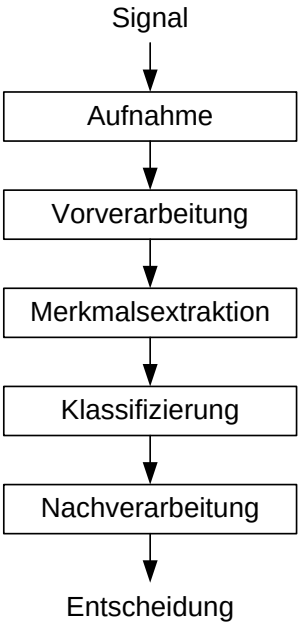


Abbildung 2.1: Aufbau eines Systems zur Mustererkennung.

- Bekräftigungslernen: Der Mustererkenner bekommt eine Rückmeldung über gemachte Klassifikationen in Form von Wahr oder Falsch. Eine Aussage über die tatsächlich vorliegende Klassenzugehörigkeit wird nicht getroffen.

2.1.2 Arbeitsphase

Zu Beginn wird ein Signal mit Hilfe eines Sensors, beispielsweise einer Kamera oder einem Mikrophon aufgenommen. Diese digitalen Daten werden vorverarbeitet, indem die Muster in kleinere Segmente zerlegt werden. Zudem kann eine Reduktion von irrelevanten Signalanteilen sowie eine Vereinheitlichung des Wertebereichs durchgeführt werden.

Nach der Vorverarbeitung kommt es zur Extraktion von Merkmalen, welche die Muster in den folgenden Schritten repräsentieren. Durch die Berechnung der Merkmale findet eine Reduktion der Datengröße statt. Merkmale werden anwendungsspezifisch so definiert, dass Objekte gleicher Klasse ähnliche Werte und Objekte unterschiedlicher Klassen möglichst verschiedene Werte ergeben. Die aufgestellten Merkmale sollten zudem möglichst einfach berechenbar und unempfindlich gegenüber für das Klassifikations-

problem irrelevanten Transformationen des Eingangssignals sein.

Die Aufgabe des Klassifizierers ist nun, die durch Merkmalsvektoren charakterisierten Objekte einer Klasse zuzuweisen. Die Modellparameter für die Entscheidungsfunktion des Klassifizierers werden während der Trainingsphase angelernt. Eine Überanpassung durch zu komplexe Modelle an die Trainingsdaten ist zu vermeiden. Sie äußert sich in einem, im Vergleich mit der Trainingsphase, erhöhten Klassifikationsfehler während der Arbeitsphase. Die Generalisierungsfähigkeit, also die Fähigkeit unbekannte Objekte zu klassifizieren, wird dadurch verschlechtert.

Zum Abschluss kommt es zu einer Nachverarbeitung der Klassifikationsergebnisse. Diese kann kontextabhängig erfolgen, d.h. aufeinander folgende, benachbarte Objekte fließen in die finale Entscheidung mit ein.

Die Qualität eines Mustererkenners spiegelt sich normalerweise in der Fehlerrate wieder. Abhängig von der Anwendung kann es sich jedoch herausstellen, dass sich Fehlklassifikationen einer bestimmten Klasse für die Anwendung als teurer erweisen, als Fehlklassifikationen von Objekten einer anderen Klasse. In diesem Fall kann die Nachverarbeitung dazu genutzt werden, die teuren Fehler zu vermeiden.

2.2 Räumliches Hören

Das wissenschaftliche Fachgebiet des räumlichen Hörens erforscht den Zusammenhang von akustischen Schallereignissen (=Reizgröße) und der Richtung bzw. räumlichen Ausdehnung ihrer wahrgenommenen Hörereignisse (=Empfindungsgröße).

Lokalisation und Ausdehnung der Hörereignisse können mit Hilfe des in Abb. 2.2 gezeigten kopfbezogenen Polarkoordinatensystems und seinen Koordinaten Azimut φ, Elevation δ und Entfernung r beschrieben werden. Im Folgenden soll lediglich auf für die Diplomarbeit relevanten Aspekte eingegangen werden.

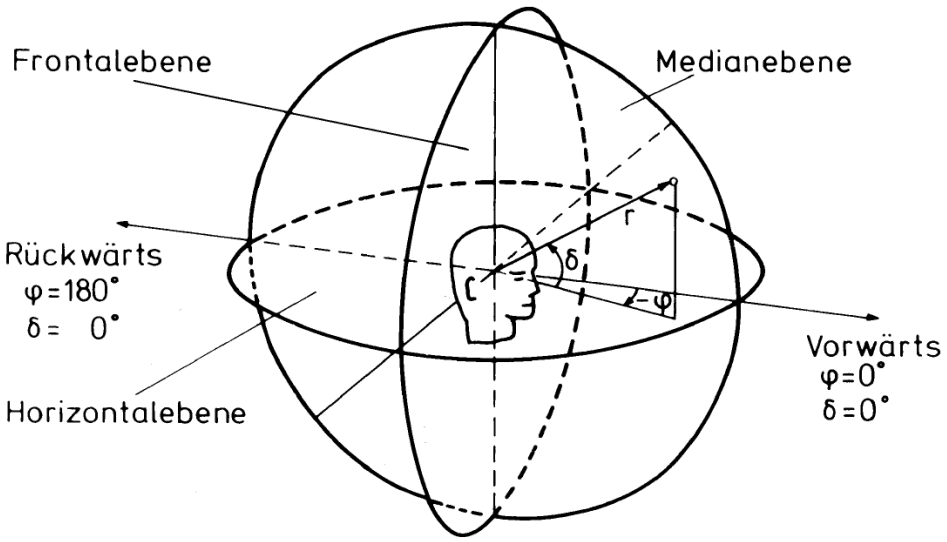


Abbildung 2.2: Kopfbezogenes Koordinatensystem. Bild aus [6].

2.2.1 Lokalisation bei einer Schallquelle

Bei Schalleinfall aus Richtung der Horizontalebene (siehe Abb. 2.2) wertet das Gehör nicht nur monaurale Ohrsignale aus. Darüber hinaus spielen interaurale Ohrsignalmerkmale, also Merkmale die auf Unterschieden zwischen den an den Trommelfellen anliegenden Schallsignalen basieren, für die Präzision der Lokalisation eine wesentliche Rolle [6].

Schallwellen mit kleiner Wellenlänge im Vergleich zu den Kopfabmessungen werden durch den Kopf abgeschattet. Der Schallpegel am der Schallquelle zugewandten Ohr weist einen höheren Wert auf als der Pegel am abgewandten Ohr. Es kommt zur sogenannten interauralen Pegeldifferenz, welche besonders für Frequenzen über ca. 1,5 kHz für die Lokalisation relevant ist. Schallwellen mit großer Wellenlänge im Vergleich zu den Kopfabmessungen hingegen, werden um den Kopf gebeugt. Das Gehör wertet die Laufzeitdifferenz zwischen des jeweiligen Eintreffens des Schalls an den Ohren aus. Diese interaurale Zeitdifferenz bestimmt die Lokalisation für Frequenzen bis ca. 1,5 kHz [7].

Die Merkmale Interaurale Zeit- und Pegeldifferenz sind nicht in der Lage, jede mögliche Richtung eines Schallereignisses eindeutig darzustellen. Es treten Mehrdeutigkeiten zwischen Schall von vorne und hinten auf. Deshalb bedarf es weiterer Merkmale um eine eindeutige Lokalisation möglich machen.

Visuelle Informationen können eine wichtigere Rolle spielen als die auditive Information, wie z.B. der Bauchredner-Effekt zeigt [8]. Ebenso können durch Kopfbewegungen bezüglich einer oder mehrerer Schallquellen weitere dynamische Informationen, die in der Regel über statisch gewonnene auditive Informationen dominieren, gewonnen werden. Ein Experiment von Blauert [9] zeigt die Bedeutung von spektralen Merkmalen, wenn interaurale Zeit- und Pegeldifferenzen näherungsweise gleich sind. Es stellte sich heraus, dass die Lokalisation von richtungsbestimmenden Bändern beeinflusst wird. Zeigt ein Ohrsignal z.B. eine starke Anhebung um 8 kHz so wird das Hörereignis von oben wahrgenommen (Elevationseffekt). Ein Schallereignis wird durch die jedem Menschen eigene Filtereigenschaft durch Kopf, Oberkörper und Ohrmuschel abhängig vom Elevationswinkel spektral verändert. Diese Änderung wird durch die Außenohr-Übertragungsfunktionen (Englisch: Head-Related Transfer Functions, HRTFs) beschrieben.

2.2.2 Lokalisation bei mehreren Schallquellen

In diesem Abschnitt sollen die psychoakustischen Effekte, die bei einer Überlagerung von mehreren Schallereignissen auftreten, besprochen werden. Diese Effekte werden bei der Lautsprecher-Wiedergabe ausgenutzt, um Hörereignisse zwischen zwei Lautsprechern zu postieren. Vorerst wird von einer Überlagerung des Schalls zweier Lautsprecher in Stereo-Standardaufstellung mit identischen oder sehr ähnlichen Signalen ausgegangen.

Die Lokalisation von Breitbandsignalen unterschiedlicher Pegel- und Zeitdifferenzen der Lautsprechersignale kann Abbildung 2.3 entnommen werden [10]. Es kommt für den abgebildeten Wertebereich von Pegel- bzw. Laufzeitdifferenzen zu einer Summenlokalisation, d.h. die Signale beider Lautsprecher werden als Phantomschallquelle, eine virtuelle Schallquelle die sich zwischen beiden Lautsprechern befindet, wahrgenommen. Allerdings ist die Lokalisation einer virtuellen Schallquelle weniger scharf als die einer realen Schallquelle. Die räumliche Abbildung der virtuellen Schallquelle wird darüber hinaus als unpräziser und in manchen Fällen diffuser als eine korrespondierende reale Schallquelle wahrgenommen [11]. Ein weiterer Effekt ist die Elevation von Phantomschallquellen [12]. Die Wahrnehmung der Lokalisation erfolgt höher als die waagerechte

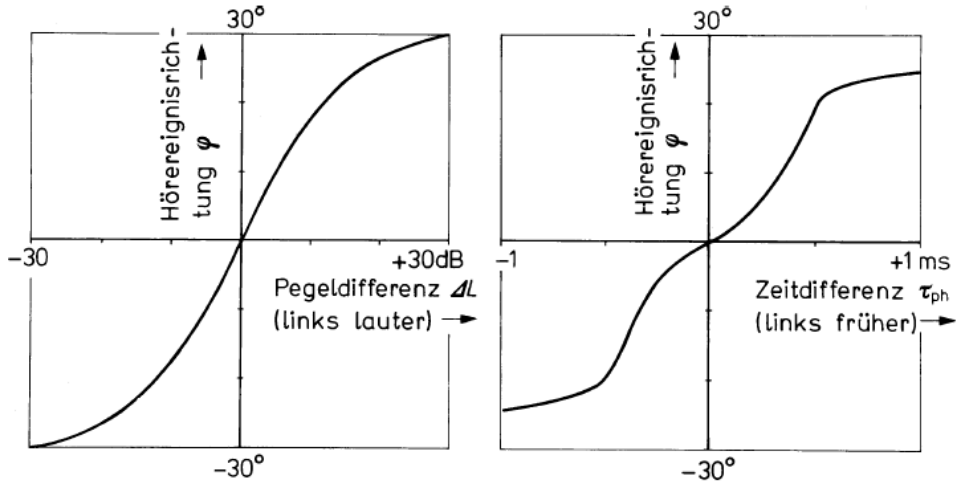


Abbildung 2.3: Summenlokalisationskurven von Breitbandsignalen. Bild aus [6].

Verbindungslinie zwischen den Lautsprechern.

Beträgt der Laufzeitunterschied zwischen beiden Lautsprechersignalen jedoch mehr als etwa 1 ms kommt es zum sogenannten Präzedenzeffekt. Das Hörereignis wird aus Richtung der zuerst eintreffenden Wellenfront wahrgenommen. Dieses Verhalten des Gehörs ermöglicht die Lokalisation von Schallereignissen in geschlossenen Räumen. Der durch Reflexionen entstehende Schall wird durch das Gehör im Hinblick auf die Bildung der Hörereignisrichtung nicht berücksichtigt. Ab einer bestimmten Verzögerungszeit, der sogenannten Echoschwelle, zerfällt das Hörereignis. Es wird zusätzlich ein Echo aus Richtung der verzögerten Wellenfront wahrgenommen. Die Echoschwelle kann abhängig von der Art des Signals zwischen 1 ms und mehreren 100 ms betragen.

2.2.3 Räumliche Ausdehnung von Hörereignissen

Für die Wahrnehmung der räumlichen Ausdehnung von Schallereignissen ist die Ähnlichkeit der an den Trommelfellen anliegenden Signale von großer Bedeutung. Als Maß für die Ähnlichkeit dient oftmals die interaurale Kohärenz [13]:

$$k = \max_{\tau} \frac{\sum_{n=-\infty}^{\infty} e_R[n] e_L[n+\tau]}{\sqrt{\sum_{n=-\infty}^{\infty} e_R^2[n] \sum_{n=-\infty}^{\infty} e_L^2[n+\tau]}} \tag{2.1}$$

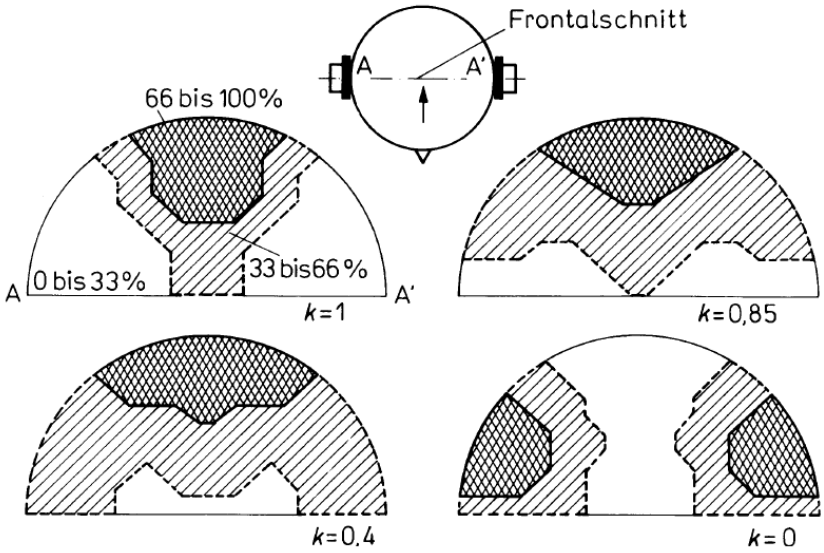


Abbildung 2.4: Einfluss der Kohärenz auf Lage und Ausdehnung von Hörereignissen. Bild aus [6].

wobei es sich bei $e_R[n]$ und $e_L[n]$ um die Ohrsignale des rechten bzw. linken Ohrs handelt. Das Maximum der normierten Kreuzkorrelationsfunktion ergibt den Kohärenzwert, welcher sich zwischen 0 für nicht kohärente und 1 für kohärente Signale bewegt.

Abbildung 2.4 veranschaulicht ein Experiment von Cherniak [14] während dessen Versuchspersonen die Lage und Ausdehnung von Hörereignissen, hervorgerufen durch Signale mit unterschiedlicher interauraler Kohärenz, angeben sollten. Die mit Prozentzahlen versehenen markierten Bereiche stellen die relative Häufigkeit dar, mit denen die Versuchspersonen den angegebenen Bereich als Hörereignis angeben hatten. Bei kohärenten Signalen wird ein Hörereignis relativ schmal in der Mitte lokalisiert. Mit sinkender Kohärenz dehnt sich das Hörereignis räumlich aus, bis es bei $k = 0$ in zwei Teile zerfällt.

2.3 3/2-Lautsprecherwiedergabe

Die weit verbreitete 3/2-Lautsprecherwiedergabe für Mehrkanalton wird im Standard ITU-R BS. 775-1 [15] geregelt. Diese Lautsprecheranordnung ist abwärtskompatibel zu

mehreren Formaten. Das bedeutet, dass z.B. Zweikanalstereo- und Monoformate ohne eine Änderung der Lautsprecheraufstellung abgespielt werden können. Wie in Abbildung 2.5 gezeigt, wird die Zweikanalstereo-Anordnung mit ihren Lautsprechern L und R um einen Center-Kanal C erweitert. Dies bewirkt, dass mittige Hörereignisse auch in nicht idealer Hörposition aus Center-Richtung wahrgenommen werden. Im Gegensatz dazu kommt es bei Wiedergabe über die Zweikanalstereo-Anordnung in nicht idealer Hörposition aufgrund von Laufzeitunterschieden und daraus resultierendem Präzedenzeffekt (siehe Kapitel 2.2.2), zu einer Verschiebung der mittigen Hörereignisse in Richtung des Lautsprechers in unmittelbarer Nähe des Hörers.

Die Mitte des ITU-Abhörkreises stellt die optimale Sitzposition dar, während die äußeren vier in der Abbildung eingezeichneten Sitzpositionen die ungünstigsten Sitzpositionen markieren.

Mit den zwei Lautsprechern LS und RS ist es nicht möglich Phantomschallquellen in einem Azimut-Winkelbereich von 30° bis 330°, im seitlichen und hinteren Bereich, zwischen den Lautsprechern stabil zu positionieren [16]. Dies gelingt nur im vorderen Bereich aufgespannt durch die Lautsprecher L, C und R.

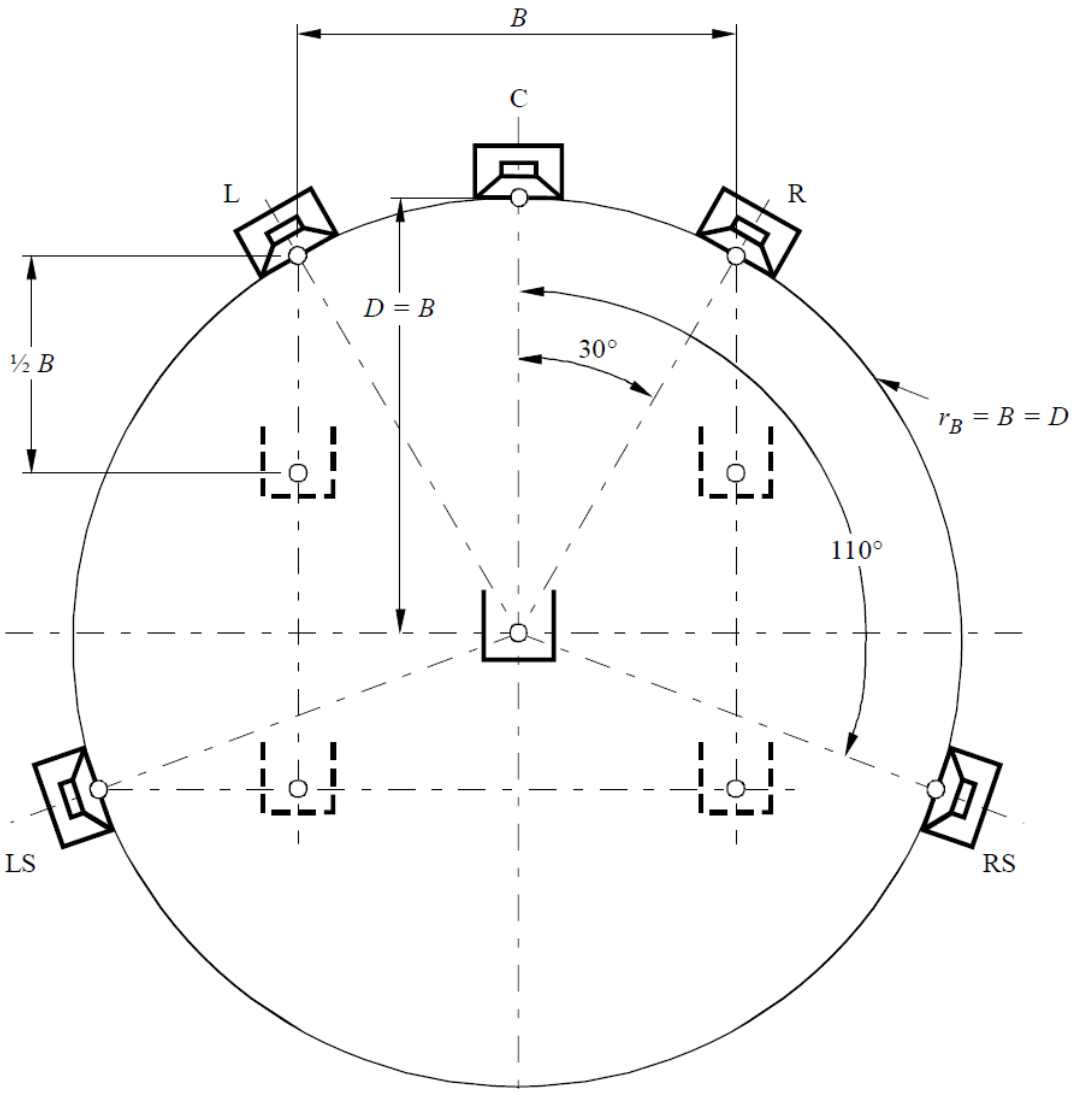


Abbildung 2.5: ITU Abhörkreis mit 3/2-Lautsprecheranordnung. Bild aus [17]

3 Stand der Technik

Dieses Kapitel soll den Stand der Technik, auf dem die im Rahmen der Arbeit durchgeführten Erweiterungen basieren, beschreiben. In Kapitel 3.1 wird der Aufbau und die Funktionsweise des Systems zur Sprachdetektion beschrieben. Methoden zur Mehrkanalkanaltonerweiterung werden in Kapitel 3.2 besprochen.

3.1 Sprachdetektion

Die Sprachdetektion (Englisch: Voice Activity Detection, VAD) beschäftigt sich mit dem Identifizieren von Segmenten innerhalb eines Signals, in denen Sprache enthalten ist. Anwendung findet die Sprachdetektion u.a. in den Bereichen Sprachkodierung [18], Sprachverbesserung [19] und Spracherkennung [20] als Vorverarbeitung zur Lokalisation von Sprache innerhalb eines Signals [21].

Das hier beschriebene System zur Sprachdetekion ist eine Anwendung der Mustererkennung, die in Kapitel 2.1 beschrieben wurde. Die grundsätzliche Funktionsweise der Sprachdetekion kann Abbildung 3.1 entnommen werden.

Die Merkmalsberechnung ist ein wichtiger Schritt, um die im Signal $x[n]$ enthaltenen

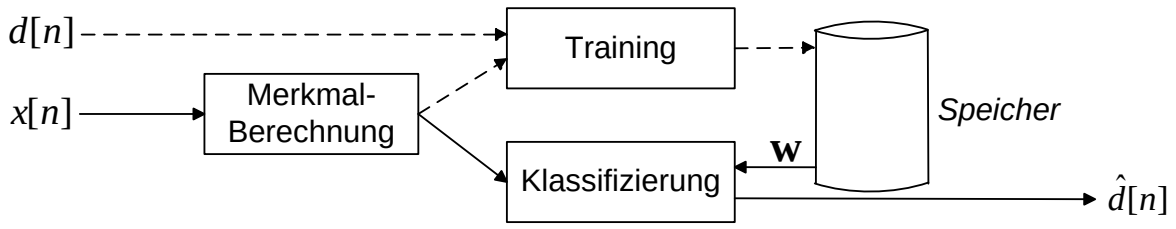


Abbildung 3.1: Überblick über die Funktionsweise der Sprachdetektion. Bild aus [22].

Muster der Sprache in Form von Zahlenwerten erfassen und beschreiben zu können. Da man im Vorfeld weiß, dass die Anwendung in der Detektion von Sprache liegt, fließt also bei der Definition der Merkmale Vorwissen über die zu detektierende Klasse und das Signal mit ein [4]. Die im Rahmen der Arbeit verwendeten Merkmale werden in Kapitel 3.1.2 beschrieben.

Die Aufgabe eines Klassifizierers besteht darin, durch Merkmale charakterisierte Objekte, einer Klasse zuzuordnen. Der im Rahmen der Arbeit verwendete Klassifizierer, das künstliche neurale Netzwerk, wird in Kapitel 3.1.1 beschrieben.

Da es sich bei der Sprachdetektion um eine komplexe Anwendung der Mustererkennung handelt, wird der Klassifizierer vor der eigentlichen Anwendung mit Hilfe von Trainingsdaten angelernt. Während der Trainingsphase, im Blockschaltbild veranschaulicht durch strichlierte Pfeile, ist die Klassenzugehörigkeit $d[n]$ jedes Merkmalvektors für den Klassifizierer sichtbar. Dadurch ist er in der Lage seine inneren Modellparameter $\mathbf{W}$ auf die gewünschte Separation der Klassen einzustellen. Die trainierten Modellparameter werden gespeichert für die darauf folgende Testphase. Nun werden dem Klassifizierer Testdaten zugeführt, deren Klassenzugehörigkeit nicht bekannt ist. Diese soll durch den trainierten Klassifizierer geschätzt werden.

3.1.1 Klassifizierer

Im Bereich der Sprachdetektion kommen zahlreiche Klassifizierer zum Einsatz. Einige Beispiele hierfür sind u.a. der auf Fisher's Kriterium basierende Klassifizierer [4], Klassifizierer basierend auf Gaußschen Mischverteilungen [23] oder künstliche neurale Netzwerke [24]. Die genannten Klassifizierer wurden in [19] für die Detektion von Sprache in Filmen miteinander verglichen. Das neuronale Netzwerk lieferte dabei unter den getesteten Klassifizierern die beste Erkennungsrate. Deshalb wurde das in [19] verwendete künstliche neurale Netzwerk der Netlab Toolbox [25] im Rahmen dieser Arbeit verwendet.

Künstliches neurales Netzwerk

Die künstlichen neuralen Netzwerke (KNN) [24] sind biologischen neuronalen Netzwerken, wie sie zum Beispiel im zentralen Nervensystem des Menschen vorkommen, nachempfunden. Der prinzipielle Aufbau kann Abbildung 3.2 entnommen werden. Im Rahmen der Arbeit wird eine einfache Klasse der KNNs verwendet, das Multi-Layer Perceptron (MLP). Zwischen Ein- und Ausgangsschicht eines MLPs befinden sich zusätzlich eine- oder mehrere verdeckte Schichten. Die Verbindung zwischen den einzelnen Neuronen eines Netzwerkes werden hergestellt mit Synapsen. Der Ausgang $y_j[n]$ eines Neurons der Schicht j feuert, wenn die Summe der gewichteten Eingangssignale $v_j[n]$ einen gewissen Schwellwert überschreiten. Dieses Verhalten wird durch eine nichtlineare differenzierbare Aktivierungsfunktion dargestellt. Oft wird hierfür eine sigmoide Funktion verwendet:

$$\varphi(v_j[n]) = \frac{1}{1 + e^{-v_j[n]}} \tag{3.1}$$

wobei sich das Eingangssignal $v_j[n]$ der Aktivierungsfunktion φ aus der gewichteten Summe aller Synapsen plus Biaswert $\mathbf{w}_{j0}[n]$ eines Neurons der Schicht j berechnet. Der Index n steht für die n-te Iteration der Gewichte des KNNs während der Trainingsphase.

$$v_j[n] = \sum_{i=0}^{m} \mathbf{w}_{ji}[n] y_i[n] \tag{3.2}$$

Der Index i bezieht sich dabei auf die Neuronen der vorhergehenden Schicht. Der Ausgang $y_j[n]$ eines Neurons zur Iteration n errechnet sich somit zu:

$$y_j[n] = \varphi_j(\sum_{i=0}^{m} \mathbf{w}_{ji}[n] y_i[n]) \tag{3.3}$$

Die Gewichte $\mathbf{W}$ des Netzwerkes werden während der Trainingsphase, über die Lösung eines Optimierungsproblems einer definierten Kostenfunktion eingestellt.

Kostenfunktion

Der Fehler $e_j[n]$ eines Neurons j zum Zeitpunkt der Iteration n ergibt sich aus der Klasse der manuell gekennzeichneten Referenz $d[n]$ und durch den Ausgang des Neurons j in der Ausgangsschicht:

$$e_j[n] = d_j[n] - y_j[n] \tag{3.4}$$

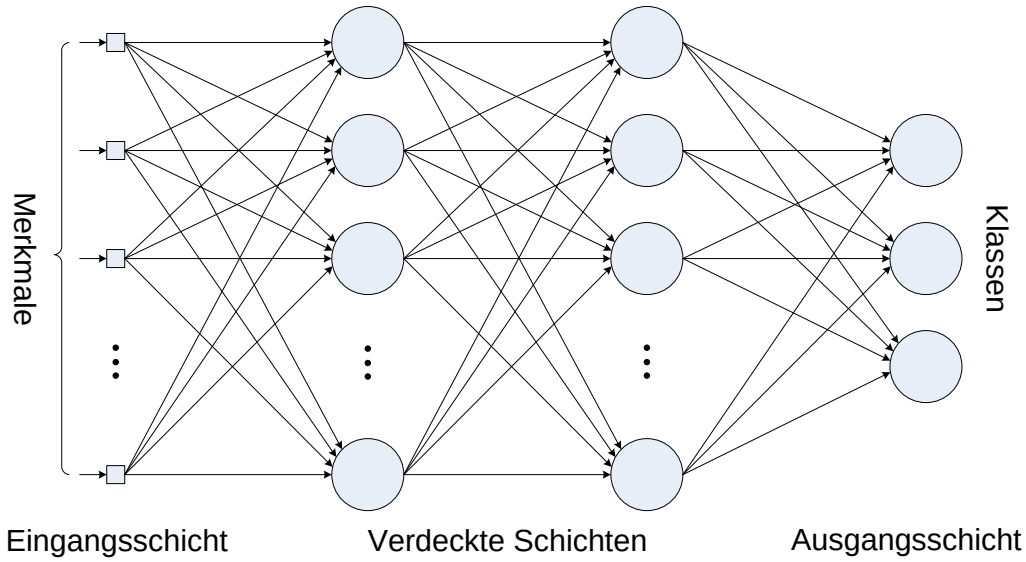


Abbildung 3.2: Der prinzipielle Aufbau eines künstlichen neuronalen Netzwerks. Bild aus [24].

Die momentane Energie des Fehlers eines Neurons ist definiert als $\frac{1}{2}e_j^2[n]$. Die Gesamtenergie E des Fehlers über alle Neuronen der Ausgangsschicht C lautet:

$$E[n] = \frac{1}{2}\sum_{j \in C} e_j^2[n] \tag{3.5}$$

Die mittlere Energie des Fehlers $E_{av}[n]$ mit

$$E_{av}[n] = \frac{1}{N}\sum_{n=1}^{N} E[n] \tag{3.6}$$

über alle N Trainingsmuster ist in der Lage, die Qualität des Lernverhaltens des Netzwerkes zu beschreiben. Die sogenannte Kostenfunktion $E_{av}[n]$ kann nun durch Lernalgorithmen, beispielsweise durch ein Gradientenverfahren minimiert werden [24] mit dem Ziel, die für die Anwendung optimalen Gewichtungsparameter $\mathbf{W}$ des Netzwerkes zu bestimmen.

3.1.2 Berechnung der Merkmale

Die Berechnung von Merkmalen dient dem Zweck, Eigenschaften von Mustern eines Signals mit Hilfe von Zahlenwerten beschreiben zu können (siehe Kapitel 2.1) um sie in

Folge einer Klasse zuweisen zu können. In diesem Kapitel soll hauptsächlich auf die im Rahmen der Arbeit verwendeten Merkmale eingegangen werden.

Merkmale können sich u.a. auf Eigenschaften des Signals im Zeit- oder im Frequenzbereich beziehen. Um ein digitales Signal in den Frequenzbereich zu transformieren, wird die Kurzzeit-Fouriertransformation (Englisch: Short-time Fourier Transformation, STFT) [26] verwendet:

$$\mathrm{STFT}\{x[n]\} = X[m,k] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jkn} \tag{3.7}$$

Da Signale $x[n]$ mit digitaler Zeitbasis n verwendet werden, handelt es sich bei Gleichung 3.7 um eine diskrete STFT. Bei m und k handelt es sich um den Zeit- bzw. Frequenzindex der STFT. Die charakteristische Eigenschaft der STFT ist die Anwendung einer Fourier Transformation auf einen begrenzten Ausschnitt des Signals $x[n]$ durch eine Multiplikation mit der Fensterfunktion $w[n]$. Durch die Fensterung werden Signal $x[n]$ in periodischen Abständen (Zeitindex m) Teile entnommen, die einer Fouriertransformation unterzogen werden. Da es sich bei einem Audiosignal üblicherweise um kein stationäres Signal handelt, können so Veränderungen im Frequenzbereich über die Zeit erfasst werden. Während dieser Arbeit wurde ein Hann Fenster mit einer Überlappung von 50% sowie einer Fensterbreite von etwa 23 ms verwendet. In diesem Zeitbereich kann Sprache als quasistationär angenommen werden. Letztendlich steht also alle 11.5 ms ein neues Frequenzspektrum $X[m,k]$ zur Verfügung.

Spectral Centroid

Der Spectral Centroid (SC) beschreibt die Lage des Massenschwerpunkts eines Spektrums und wird über das normalisierte Amplitudenspektrum $\widetilde{X}[k,m]$ berechnet [27]:

$$\widetilde{X}[k,m] = \frac{|X[m,k]|}{\sum_{k\in\kappa_+}|X[m,k]|} \tag{3.8}$$

Die Summation erfolgt über alle Frequenzindizes k, die Teil der Menge der nicht-negativen Frequenzindizes κ_+ sind.

$$\mathrm{SC}[m] = \sum_{k\in\kappa_+} k\widetilde{X}[m,k] \tag{3.9}$$

In einer alternativen Definition des Spectral Centroids [28] kann die Berechnung auch mit Hilfe der spektralen Leistungsdichte erfolgen, die entweder über das gemittelte Periodogramm [29] oder die Fouriertransformierte der Autokorrelationsfunktion des Signals ermittelt wird.

Spectral Spread

Der Spectral Spread (SSP) [27] ist das zweite zentrale Moment des Spektrums und beschreibt die Ausdehnung des Spektrums in Bezug auf den Schwerpunkt SC. Die Berechnung von $\widetilde{X}[m,k]$ erfolgt nach Gleichung 3.8.

$$\mathrm{SSP[m]} = \sum_{k \in \kappa_+} (k - \mathrm{SC}[m])^2 \widetilde{X}[m,k] \tag{3.10}$$

Der SSP unterscheidet sich für Spektren eines Signals die eher tonal (geringe Ausdehnung des Spektrums) oder rauschartig (große Ausbreitung des Spektrums) sind. Eine alternative, im Rahmen des MPEG-7 Standards [28] verwendete Definition des SSPs, verwendet eine logarithmische Frequenzeinteilung und wird über die spektrale Leistungsdichte berechnet.

Spectral Flux

Der Spectral Flux (SF) [30] beschreibt den Unterschied zweier zeitlich aufeinander folgenden Spektren. Er wird für jedes Frequenzband $[l_r; u_r]$ mit Subbandindex r, wobei $1 \leq r \leq 4$ berechnet. l_r und u_r stehen hierbei für die untere und obere Grenze eines Subbands.

$$\mathrm{SF}[m,r] = \sqrt{\sum_{k=l_r}^{u_r} \left(|X[m,k]| - |X[m-1,k]|\right)^2} \tag{3.11}$$

Als Maß für den Unterschied zwischen zwei Spektren wird die Euklidische Distanz verwendet.

Spectral Flatness Measure

Das Spectral Flatness Measure (SFM) [31] charakterisiert die Abweichung eines Spektrums von einem flachen Spektrum. Es ist definiert als Quotient aus geometrischem und arithmetischem Mittel.

$$\mathrm{SFM}[m,r] = \frac{e^{\left(\sum_{k=l_r}^{u_r} \log\left(|X[m,k]|\right)\right)/L}}{\frac{1}{L}\sum_{k=l_r}^{u_r} |X[m,k]|} \tag{3.12}$$

L steht für die Anzahl der spektralen Koeffizienten. Die Ungleichung vom arithmetischen und geometrischen Mittel besagt, dass das arithmetische Mittel stets größer oder gleich groß wie das geometrische Mittel ist. Daraus ergibt sich, dass SFM nie größer als eins werden kann. Für weißes Rauschen wird das SFM zu eins. Für weniger flache Spektren liegt der Wert unterhalb von eins.

Relative Spectral Perceptual Linear Prediction

Das Relative Spectral Perceptual Linear Prediction Merkmal (RASTA-PLP) [32] wird für jeden STFT-Frame folgendermaßen berechnet [19]:

1. Kompression der Magnitude der spektralen Koffizienten des spektralen Leistungsdichte.
2. Bandpassfilterung des Zeitverlaufs der spektralen Koeffizienten.
3. Expansion der Amplitude der spektralen Koeffizienten.
4. Multiplikation mit Kurve gleicher Lautstärke und Potenzieren mit Exponent 0.33 um die Lautstärkewahrnehmung des Menschen zu simulieren.
5. Berechnung eines All-Pole Modells des Spektrums.

Perceptual Linear Prediction

Die Berechnung der Perceptual Linear Prediction Merkmale (PLP) [33] erfolgt analog zu den RASTA-PLP Merkmalen, wobei die Schritte 1-3 allerdings nicht ausgeführt werden.

Mel Frequency Cepstral Coefficients

Für die Berechnung der Mel Frequency Cepstral Coefficients (MFCC) [34] wird das Spektrum mit Dreiecksfiltern in Bänder zusammengefasst. Die Bandbreite der Filter richtet sich nach der Melskala [35] und wächst mit steigender Frequenz. Die Beträge der Bänder werden logarithmiert und mit Hilfe einer diskreten Cosinustransformation [36] dargestellt. Die errechneten Koeffizienten der Transformation entsprechen den MFCCs.

Panning Index Centroid

Der Panning Index Centroid ist ein effiziente Berechnung des Panning Indexes (siehe Kapitel 3.2.3) [37], der die Position von amplitudengepannten Signalen im Stereobild ermittelt. Der Wertebereich des Merkmals erstreckt sich von -1 für links gepannte Quellen bis 1 für rechts gepannte Quellen.

$$PI2[m,k] = \frac{X_R[m,k] - X_L[m,k]}{X_R[m,k] + X_L[m,k]} \tag{3.13}$$

3.1.3 Nachverarbeitung der Merkmale

Delta Merkmale

Delta Merkmale beschreiben die Änderung eines Merkmales über die Zeit. Es handelt sich hierbei um die zeitliche Ableitung der Merkmale, die oft über eine Faltung mit einer linearen Rampe berechnet wird. Eine effiziente Methode ist auch die Filterung mit einem IIR-Hochpassfilter.

Sigma Merkmale

Während Delta Merkmale die Änderung eines Merkmales über verhältnismäßig kurze Zeit beschreiben, sind Sigma Merkmale [22] hingegen dafür vorgesehen die Merkmalswerte über längere Zeitbereiche zu mitteln. Sigma Merkmale komplementieren die aus der STFT gewonnenen Kurzzeit-Merkmale, indem sie zusätzlich Informationen von Nachbarmerkmalen über einen längeren Zeitraum mit einbeziehen. Eine effiziente Methode

der Berechnung ist die Tiefpassfilterung einer Merkmalsfolge, ganz in Analogie zu den Delta Merkmalen.

Zentrierung und Varianznormalisierung

Nach der Ermittlung aller Merkmale werden diese normiert und somit in einen einheitlichen Wertebereich transformiert. Es wird hierbei der Mittelwert jedes Merkmals entfernt sowie die Varianz normalisiert indem man jedes Merkmal durch seine Standardabweichung dividiert. In der Testphase werden die Merkmale mit den aus der Trainingsphase ermittelten Werten für Mittelwert und Varianz zentriert bzw. varianznormalisiert.

3.1.4 Merkmalsselektion

Die Selektion von Merkmalen dient der in erster Linie der Verbesserung der Klassifikationsgenauigkeit. Ebenso werden der benötigte Speicherplatz sowie der Berechnungsaufwand reduziert. Dies schlägt sich nicht zuletzt in einem geringeren Zeitaufwand für das Training des Klassifizierers nieder [38].

In den folgenden Abschnitten wird auf verschiedene Ansätze zur Merkmalsselektion eingegangen.

Filtermethoden

Filtermethoden zur Merkmalsselektion bewerten einzelne Merkmale oder Untermengen von Merkmalen anhand eines Bewertungskriteriums. Dies geschieht unabhängig vom Klassifizierer als Vorverarbeitung. Ein etabliertes Bewertungskriterien ist das Fisher Kriterium: Es beschreibt die Überlappung der Merkmalsverteilungen von Klassen. Eine geringe Überlappung, und damit eine hohe Klassifikationsgenauigkeit, ergibt sich für eine große Distanz zwischen den Mittelpunkten der Merkmale beider Klassen bei gleichzeitiger geringer Varianz der Merkmale der einzelnen Klassen [39]. Korrelationskriterien, wie u.a. der Pearson Koeffizient [40], beschreiben den Grad der Korrelation der durch Merkmale dargestellten Objekte zur manuell erstellten Referenz der Klassen.

Das Bewertungskriterium Transinformation bewertet den Informationsgehalt, den die Merkmale der Objekte und die Klassenreferenz gemeinsam haben.

Der Nachteil der Filtermethode ist, dass die berechnete Teilmenge der Merkmale in keinem Zusammenhang zu ihrer späteren Anwendung in Verbindung mit dem Klassifizierer steht [38]. In [41] konnte gezeigt werden, dass die durch Filtermethoden bestimmten Teilmengen der Merkmale keine irrelevanten Merkmale enthalten, die jedoch die Qualität der Klassifikation erhöhen würden. Ebenso können durch Filtermethoden keine korrelierten Merkmale aus der Teilmenge entfernt werden, die das Klassifikationsergebnis verschlechtern könnten.

Wrappermethoden

Bei den Wrappermethoden findet eine Bewertung eines Merkmals oder einer Teilmenge von Merkmalen durch den Klassfizierer selbst statt. Die Qualität des Klassifikationsergebnisses kann beispielsweise mit einem Fehlermaß über eine Kreuzvalidierung [42] (siehe Kapitel 5.1.1) oder über statistische Tests [43, 44] ermittelt werden.

Für die Zusammenstellung der Teilmengen der Merkmale existieren diverse Algorithmen [41] . Ein im Rahmen der Arbeit angewendeter Algorithmus ist die sequentielle Vorwärtsselektion [45]. Hierfür werden die besten verbliebenen Merkmale schrittweise zur bestehenden Teilmenge hinzugefügt, bis keine Verbesserung der Klassifikationsgenauigkeit mehr erreicht werden kann.

Der Nachteil von Wrappermethoden kann der hohe Zeit- und Berechnungsaufwand sein [38]. Darüber hinaus besteht für eine kleine Menge von Trainingsdaten die Gefahr der Überanpassung der Merkmalauswahl an den vorliegenden Datensatz.

3.2 Mehrkanaltonerweiterung

Mehrkanaltonerweiterung (Englisch: Upmix) steht für die Erweiterung der Anzahl der Kanäle N eines Audiosignals auf M Kanäle, wobei gilt $M > N$. Bei $N > M$ handelt es sich um einen Downmix.

Ein Upmixer kann im Rahmen der Datenübertragung zum Einsatz kommen, denn die Kanalkapazität bei der Datenübertragung, sei es über das Internet oder über Rundfunk, ist begrenzt und mit Kosten verbunden. Dementsprechend ist die gleichzeitige Übertragung von Zweikanal- und Mehrkanalsignalen vergleichsweise teuer. Deshalb wird lediglich ein Zweikanalsignal übertragen, welches je nach Wiedergabesystem des Empfängers entweder direkt abgespielt oder über einen Upmixer auf die gewünschte Anzahl von Kanälen erweitert wird. So besteht einerseits der Vorteil der Rückwärtskompatibilität zu herkömmlichen Zweikanalstereo-Wiedergabegeräten sowie der effizienten Übertragung aufgrund der niedrigeren Kanalanzahl. Hier werden häufig Upmix-Verfahren mit Seiteninformationen wie Spatial Audio Coding verwendet.

Ein weiteres Anwendungsgebiet des Upmixers liegt in der Erzeugung von Mehrkanalton z.B. aus Film- und Musikinhalten, für die kein Mehrkanalton vorliegt.

Während der Produktion von Mehrkanalton und somit auch während des Prozesses der Mehrkanaltonerweiterung kann zwischen zwei Ansätzen unterschieden werden [46]:

- Direkt/Ambient-Ansatz: Schallquellen werden über die Frontlautsprecher verteilt, Raumanteile werden über alle Lautsprecher wiedergegeben.
- Umgebender Ansatz: Schallquellen und Raumanteile werden rund um den/die Hörer/in verteilt.

Im Folgenden soll auf verschieden Verfahren zur Mehrkanaltonerweiterung eingegangen werden.

3.2.1 Spatial Audio Coding

Das Ziel des Spatial Audio Coding (SAC) [47] ist die Kompression von Mehrkanalsignalen zur effizienten Nutzung von vorhandenen Ressourcen. Ein vorhandenes Mehr-

kanalsignal mit N Kanälen wird auf beliebige Anzahl E an Kanälen, beim Binaural Cue Coding (BCC) [48–50] auf ein Monosignal reduziert. Im Zuge des Downmixes werden die Räumlichkeit betreffende, binaurale Informationen wie Zeit-, Pegel- und Kohärenzunterschiede (siehe Kapitel 2.2) zwischen den einzelnen Kanälen, sogenannte Seiteninformationen, berechnet. Das kanalreduzierte Signal kann zusätzlich MP3- oder AAC-codiert [51, 52] und zusammen mit den Seiteninformationen übertragen oder gespeichert werden. Auf Decoderseite ist es möglich den Downmix mit E Kanälen direkt abzuspielen oder eine Mehrkanaltonerweiterung mit Hilfe der Seiteninformationen durchzuführen.

3.2.2 Matrix Verfahren

Der Zweck der matrixbasierenden Mehrkanalverfahren ist die Kodierung von Mehrkanalsignalen in Form von zwei Stereokanälen. Entstanden ist dies aus dem Umstand, dass zu Zeiten von Videorecorder (VCR) und analogem Radio nur zwei Übertragungskanäle verfügbar waren [53]. Bei Dolby Stereo handelt es sich um 4:2:4 Matrixverfahren für die Filmtonwiedergabe im Kino. Vor der Enkodierung liegen die vier Kanäle Links (L), Center (C), Rechts (R) und Surround (S) vor. Der Centerkanal wird im Enkoder um 3 dB gedämpft und den Frontkanälen L und R hinzugemischt. Der Surroundkanal durchläuft einen Bandpass von 150 bis 7000 Hz, wird um 3 dB gedämpft, 90° phasengedreht und jeweils auf den rechten Kanal addiert bzw. vom linken Kanal subtrahiert. Das fertig codierte Signal besteht aus den Kanälen L_t und R_t.

Bei der Dekodierung entsteht der Surroundkanal durch Subtraktion von L_t und R_t. Das auf beiden Kanälen gleichphasig enthaltene Centersignal wird dadurch ausgelöscht. Dieses wird wiederum aus einer Summierung von L_t und R_t erzeugt. Das Surroundsignal durchläuft Antialiasing und Tiefpassfilter, bevor es um etwa 20-30 ms in Bezug auf die Frontkanäle verzögert wird. Dies hat zur Folge, dass das Übersprechen zwischen Front- und Surroundkanälen weniger stark wahrgenommen wird, da die Schallereignisse durch den entstehenden Präzedenzeffekt bedingt (siehe Kapitel 2.2), aus Richtung der Frontlautsprecher wahrgenommen werden.

Bei der Dematrizierung kommt außerdem eine adaptive Matrix zum Einsatz, die unerwünschtes Übersprechen zwischen den Kanälen vermindern soll. Mit Hilfe der Signale L_t und R_t wird für festgelegte Frequenzbänder berechnet, in welchen vier Kanälen sich das Klanggeschehen abspielt. Dementsprechend findet eine Dämpfung der nicht genutzten Kanäle in Abhängigkeit von frequenzbandabhängigen Zeitkonstanten statt.

Dolby Surround und Dolby Pro Logic sind 4:2:4 Matrixverfahren für die Filmtonwiedergabe im Heimbereich. Dolby Surround unterscheidet sich von Dolby Stereo hauptsächlich in der einfacheren Dekodierung mit passiver Matrix mit daraus resultierender schlechterer Kanaltrennung. Besonders zwischen Center bzw. Surround und den äußeren Frontlautsprechern, beträgt die Kanaltrennung lediglich etwa 3-6 dB [54]. Bei Dolby Pro Logic werden hingegen, wie im Kinobereich, adaptive Matrizen verwendet. Der Tonmeister hat das enkodierte sowie dekodierte Mehrkanalsignal bereits während des Mischvorgangs zur Verfügung und kann so Problemen, die durch die Kodierung bzw. Dekodierung entstehen, gezielt entgegenwirken.

Dolby Pro Logic II unterstützt die Codierung mit wahlweiser 4:2:4-, 5:2:5- oder 6:2:6-Matrizierung [55]. Es stehen zwei getrennte Surroundkanäle (LS, RS) zur Verfügung. Mit Hilfe von spannungsgesteuerten Verstärkern wird die Lautstärke der einzelnen Kanäle bei der Dekodierung adaptiv so geregelt, dass unerwünschtes Übersprechen verringert wird. Das System eignet sich wegen der besseren Kanaltrennung auch für Musikproduktionen.

Lexicon Logic 7 ist ein 5:2:5 Matrixverfahren, welches kompatibel zu Dolby Surround, Dolby Pro Logic und Dolby Pro Logic II und diskret kodiertem 5.1 ist. Das Verfahren wird vorzugsweise für die Dekodierung von Mehrkanalsignalen herangezogen und eignet sich für Film und Musik.

Die Matrixverfahren SRS Circle Surround bzw. Circle Surround II können 5.1 Signale über eine 5:2:5 Matrizierung bzw. 6.1 und virtuelle 7.1 Signale über 7:2:7-Matrizierung verarbeiten [56].

3.2.3 Mehrkanaltonerweiterung mittels spektraler Gewichtung

Der im Rahmen der Arbeit verwendete Upmixer arbeitet nach dem Prinzip der spektralen Gewichtung, wie in diesem Kapitel beschrieben. Bei diesem Ansatz handelt es sich um eine blinde Mehrkanaltonerweiterung. Hierbei werden Audiosignale verarbeitet, welche zuvor nicht über einen Downmix aus einem Mehrkanalsignal erstellt wurden. Die Eingangssignale des Upmixers enthalten darüber hinaus keinerlei Zusatzdaten über räumliche Eigenschaften.

Grundidee des Verfahrens ist es, Anteile des Eingangssignals im Spektralbereich mittels Multiplikation mit einer Gewichtungsfunktion zu extrahieren. Das Klangbild des zweikanaligen Signals wird also in Direktschall und Raumanteile zerlegt und daraufhin unter Berücksichtigung der zur Verfügung stehenden Ausgangskanäle des Upmixers wieder zusammengefügt.

Berechnung von Merkmalen

Faller [57] benutzt die normalisierte Interkanal-Kreuzkorrelation für Schätzung des Direktschalls und der Raumanteile. Avendano [58–61] benutzt den Panning Index für die Beschreibung der Panoramisierung der Schallquellen, die Interkanal-Kohärenz für die Identifikation ambienter Signalanteile. Auf beide Größen wird im Folgenden näher eingegangen, da sie für Teile der im Rahmen der Diplomarbeit entstandenen Algorithmen eine Rolle spielen.

Interkanal-Kohärenz: Die Kohärenz zwischen linkem und rechten Kanal eines Zweikanalstereosignals ermöglicht es dekorrelierte, ambiente Anteile zu identifizieren.

Die Kreuzkorrelation zweier Kanäle ergibt sich aus dem Erwartungswert $\mathbf{E}$ der Multiplikation eines Frequenzbins des linken Kanals $X_l(m,k)$ mit dem konjugiert komplexen Frequenzbin des rechten Kanals $X_r(m,k)$.

$$\phi_{lr}(k) = \mathbf{E}\left\{X_l(m,k)X_r^*(m,k)\right\} \tag{3.14}$$

Die Mittelung durch den Erwartungswert erfolgt über alle Zeitframes m. Ein Audiosignal kann aber auf seine gesamte Länge bezogen im Allgemeinen nicht als stationär betrachtet

werden. Um also Veränderungen innerhalb kleinerer Zeitabschnitte zu erfassen wird eine Kurzzeit-Kreuzkorrelationsfunktion definiert:

$$\phi_{lr}(m,k) = (1-\lambda)\phi_{lr}(m-1,k) + \lambda X_l(m,k)X_r^*(m,k) \tag{3.15}$$

Diese bezieht den zuvor berechneten Kreuzkorrelationswert mit in die Berechnung ein. Wie stark vergangene und aktuelle Werte dabei einfließen bestimmt Parameter $0 \leq \lambda \leq 1$. Je kleiner λ ist, desto mehr beruht die Berechnung auf bereits vergangenen Kreuzkorrelationswerten. Mit Hilfe der Kurzzeit-Kreuzkorrelation kann die Interkanal-Kreuzkorrelation definiert werden:

$$\phi(m,k) = \frac{|\phi_{lr}(m,k)|}{\sqrt{\phi_{ll}(m,k)\phi_{rr}(m,k)}} \tag{3.16}$$

Hierbei wird der Betrag der Kreuzkorrelation durch die Wurzel der Energien des linken bzw. rechten Kanals normiert. Die Werte der Interkanal-Kohärenz bewegen sich zwischen 0 und 1, wobei stark dekorrelierte Signalanteile niedrige Kohärenzwerte ergeben.

Panning Index: Der Panning Index [61] beschreibt die Position eines Frequenzbinpaares im Stereobild zu jeder Zeiteinheit m. Mit Hilfe Gl. (3.15) definiert sich die Ähnlichkeitsfunktion ψ.

$$\psi(m,k) = 2\,\frac{|\psi_{lr}(m,k)|}{|\psi_{ll}(m,k) + \psi_{rr}(m,k)|} \tag{3.17}$$

wobei $\psi_{lr}(m,k) = \phi_{lr}(m,k)|_{\lambda=1} = X_l(m,k)X_r^*(m,k)$. Aus den partiellen Ähnlichkeitsfunktionen $\psi_l(m,k)$ bzw. $\psi_r(m,k)$ sowie der Differenzfunktion $\Delta(m,k)$ ergibt sich die Ambiguitätsfunktion $\hat{\Delta}(m,k)$, die eine Aussage darüber trifft, ob ein Frequenzbinpaar rechts oder links der Stereobildmitte positioniert ist.

$$\Delta(m,k) = \psi_l(m,k) - \psi_r(m,k) = \frac{|\psi_{lr}(m,k)|}{\psi_{ll}(m,k)} - \frac{|\psi_{lr}(m,k)|}{\psi_{rr}(m,k)} \tag{3.18}$$

$$\hat{\Delta}(m,k) = \begin{cases} 1 & falls \quad \Delta(m,k) > 0 \\ 0 & falls \quad \Delta(m,k) = 0 \\ -1 & falls \quad \Delta(m,k) < 0 \end{cases} \tag{3.19}$$

Der Panning Index Ψ ist definiert als:

$$\Psi(m,k) = (1 - \psi(m,k))\,\hat{\Delta}(m,k) \tag{3.20}$$

Dieses Merkmal kann lediglich die Richtung von Quellen, deren Panoramisierung mittels Interkanal-Pegeldifferenzen vorgenommen wurde, zuverlässig bestimmen. Wurden Aufnahmen beispielsweise nicht mit Koinzidenzmikrofonen erstellt, kommt es zu Laufzeitunterschieden, die durch den Panning Index nicht darstellbar sind.

Spektrale Gewichtung

Nachdem die Raumanteile und die Panoramisierung der Schallquellen durch die Merkmale $\zeta[k,m]$ identifiziert wurden, kommt es zur Extraktion dieser Signalanteile mittels Gewichtungsfunktion $\Gamma(\zeta[k,m])$:

$$\hat{X}[k,m] = X[k,m]\,\Gamma(\zeta[k,m]) \tag{3.21}$$

Ambienter Anteil: Wird beispielsweise ein Direkt/Ambient-Ansatz zur Mehrkanaltonerweiterung angestrebt, werden ambiente Anteile auf die Surroundkanäle geleitet. Die Gewichtungsfunktion $\Gamma(\zeta[k,m])$ soll Frequenzen, die eine weniger ausgeprägte Ambience-Charakteristik besitzen, abdämpfen. Als Merkmal $\zeta[k,m]$ für die Berechnung der spektralen Gewichte kann die Interkanal-Kohärenz dienen. Die extrahierten ambienten Anteile werden, ebenso wie bei den Matrixverfahren, mit Hilfe eines Delays zeitlich verzögert, um unter Nutzung des Präzedenzeffekts (siehe Kapitel 2.2.2) unerwünschte Lokalisationseffekte zu vermeiden. Diese können darüber hinaus durch Dekorrelation des Surroundsignals, beispielsweise mit Hilfe von Allpassfiltern [61], unterdrückt werden.

Direktschall: Bei einem Direkt/Ambient-Ansatz wird der Direktschall des zweikanaligen Eingangssignals von mindestens drei Frontlautsprechern dargestellt. Die spektralen Gewichte lassen sich über den Panning Index errechnen. Es findet hierbei eine Repanoramisierung statt, die mit Hilfe von Panning Funktionen [62–64] wie das Sinus-Gesetz [65] oder das Tangens-Gesetz [66] realisiert werden kann. Eine weitere Möglichkeit für die

Panoramisierung bei beliebigen Lautsprecheraufstellungen ist das Vector Base Amplitude Panning (VBAP) [11, 67, 68].

4 Eigenes Verfahren

4.1 Sprachdetektion

Die Inhalte diese Kapitels beziehen sich auf die eigenen Verfahren innerhalb des Moduls der Sprachdetektion. Einen Überblick über die Funktionsweise der Sprachdetektion gibt Abschnitt 3.1.

In Kapitel 4.1.1 wird eine Vorverarbeitung des Signals, noch vor der eigentlichen Berechnung der Merkmale, beschrieben. Signalanteile niedriger Kohärenz und nicht- mittig gepannten Schallquellen sollen hier gedämpft werden mit der Absicht, das Klassifikationsresultat der Sprachdetektion zu verbessern.

In Kapitel 4.1.2 werden Stereomerkmale definiert. Die Annahme ist hierbei, dass Sprache ein mittiges Panning sowie eine starke Kohärenz zwischen linkem und rechtem Kanal besitzt. Diese Eigenschaften werden mittels Interkanal-Pegeldifferenz aus Kapitel 3.2.3 und Interkanal-Kohärenz 3.2.3 beschrieben.

Eigene Verfahren für die Nachverarbeitung der durch den Klassifizierer geschätzten Klassen werden in Kapitel 4.1.3 behandelt. Einerseits wurde versucht, das Klassifikationsergebnis durch einen Klassifizierer, der während der Laufzeit eines Films Daten über Sprecher, Atmosphären, Musik usw. sammelt und auswertet, zu verbessern. Andererseits war die Verbesserung der zeitliche Positionierung von Segmentgrenzen ein Thema das bearbeitet wurde.

4.1.1 Spektrale Gewichtung

Mit Hilfe einer spektralen Gewichtung können für die Sprachdetektion irrelevante Anteile eines Signals unterdrückt werden, bevor sie durch die Mustererkennung auf die Existenz von Sprache hin untersucht werden. Es werden folgende Annahmen getroffen:

1. Sprache ist in der Mitte des Stereopanoramas positioniert.

2. Für Sprache besteht eine hohe Kohärenz zwischen linkem und rechtem Stereokanal.

Abbildung 4.1 zeigt die Funktionsweise der spektralen Gewichtung. Die in Kapitel 3.2.3 und 3.2.3 besprochenen Merkmale Interkanal-Kohärenz und Interkanal-Pegeldifferenz liefern eine räumliche Charakterisierung eines Stereosignals, was die Positionen der vorhandenen Schallquellen und ihrer Kohärenzeigenschaften zwischen den beiden Kanälen angeht. Durch eine Multiplikation geeigneter Gewichtungsfunktionen für Interkanal-Kohärenz $\Gamma_\phi(\phi[m,k])$ und Panning Index $\Gamma_\Psi(\Psi[m,k])$ mit dem Audiosignal im Frequenzbereich können unerwünschte Anteile unterdrückt werden.

$$X_{\Gamma_\phi}[m,k] = X[m,k]\,\Gamma_\phi(\phi[m,k]) \tag{4.1a}$$

$$X_{\Gamma_\Psi}[m,k] = X[m,k]\,\Gamma_\Psi(\Psi[m,k]) \tag{4.1b}$$

Spektrale Gewichtung mittels Interkanal-Kohärenz

Ambiente Signalanteile, also Anteile diffuser Natur, oder auch Rauschen besitzen kleine Interkanal-Kohärenzwerte $\phi[m,k]$ (siehe Gleichung 3.16). Diese Anteile sollen mittels einer geeigneten Gewichtungsfunktion unterdrückt werden. Es wird hierzu eine Funktion $\Gamma_\phi(\phi)$, wie sie Avendano [61] für die Ambience Extraktion während der blinden Mehrkanaltonerweiterung vorschlägt, verwendet:

$$\Gamma_\phi(\phi[m,k]) = \frac{\mu_1 - \mu_0}{2}\tanh\left(\sigma\pi(\phi[m,k] - \phi_0)\right) + \frac{\mu_1 + \mu_0}{2} \tag{4.2}$$

Wie in Abb. 4.2 zu sehen ist können Schwellwert ϕ_0 und Steilheit σ der Funktion $\Gamma_\phi(\phi)$ variabel eingestellt werden. Die Parameter μ_1 und μ_0 stehen für die oberen bzw. unteren Grenzen des Wertebereichs der Gewichtungsfunktion und lauten die Abbildung $\mu_0 = 0.1$ und $\mu_1 = 1$.

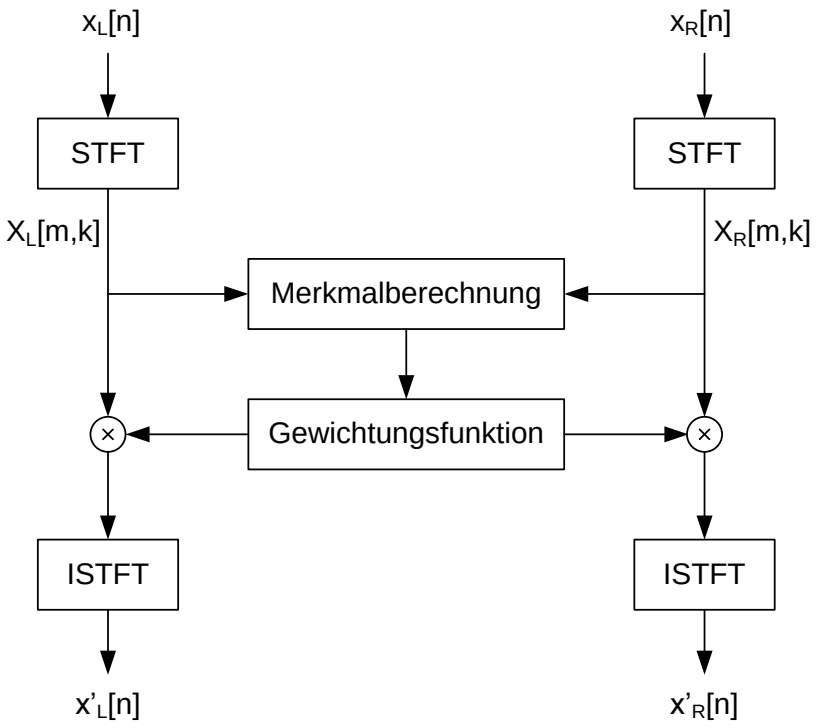


Abbildung 4.1: Funktionsweise der spektralen Gewichtung.

Spektrale Gewichtung mittels Panning Index

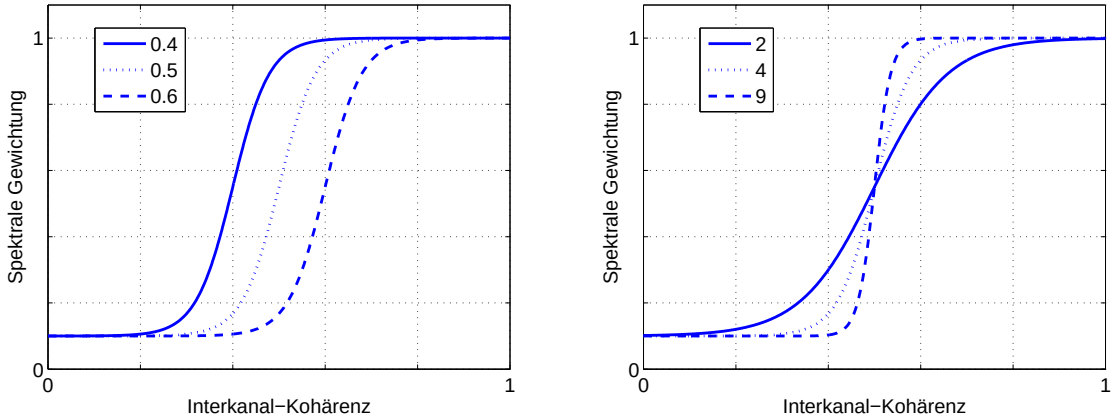
Unter Berücksichtigung der zuvor gemachten Annahmen sollen Schallquellen, die stark nach außen gepannt sind, unterdrückt werden. Die Berechnung der in Abbildung 4.1.1 veranschaulichten Gewichtungsfunktion erfolgt mit Hilfe des Panning Indexes $\Psi[m,k]$ gemäß Gl. (3.20) und basiert im Wesentlichen auf Gl. (4.2):

$$\Gamma_{\Psi}(\Psi[m,k]) = \frac{\mu_1 - \mu_0}{2} \tanh\left(\sigma\pi(\Psi[m,k] - |\Psi_0|)\right) + \frac{\mu_1 + \mu_0}{2} \tag{4.3}$$

Die Parameter μ_0, μ_1, σ und Ψ_0 der Gewichtungsfunktion verhalten sich ebenso wie für Gl. (4.2) beschrieben.

4.1.2 Stereomerkmale

Die hier aufgestellten Stereomerkmale bedienen sich der in Kapitel 3.2.3 und 3.2.3 behandelten Größen Interkanal-Pegeldifferenz und Interkanal-Kohärenz. Stereomerkmale



(a) Veränderung der Schwellwertes ϕ_0

(b) Veränderung der Steilheit σ

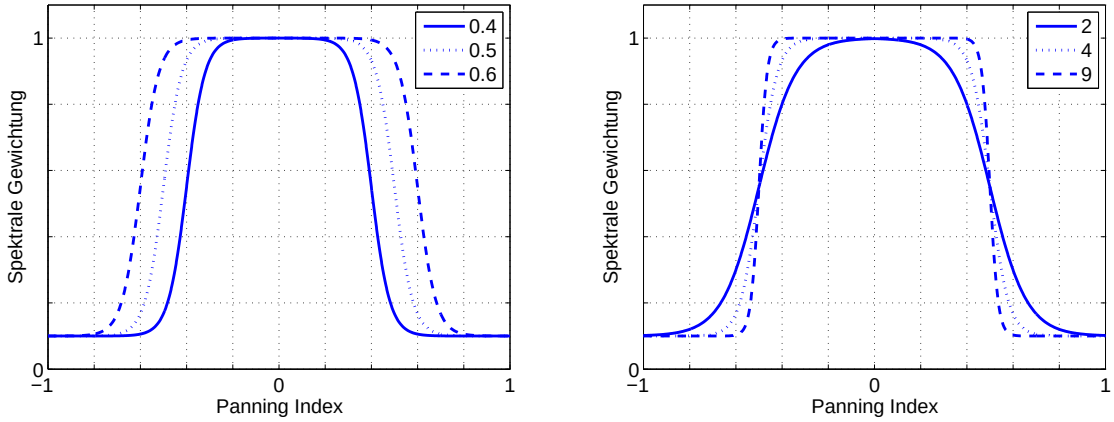
Abbildung 4.2: Interkanal-Kohärenz Gewichtungsfunktion $\Gamma_\phi(\phi)$

nutzen Informationen zwischen zwei Kanälen um das Stereoklangbild zu beschreiben. Sowohl Interkanal-Pegeldifferenz als auch Interkanal-Kohärenz sind essentielle Beschreibungsgrößen, die die Grundlage des Upmixprozesses bilden. Da diese Werte also ohnehin für den Upmixer gebraucht werden, ist nur eine einmalige Berechnung für Sprachdetektion und Upmixer vonnöten.

Panning Index Stereomerkmale

Mit Hilfe der Interkanal-Pegeldifferenz kann eine Aussage über die Lokalisierung von Schallquellen innerhalb des Stereo Panoramas getroffen werden. Eine mittig positionierte Schallquelle weist einen Panning Index von $\Psi[m,k] = 0$ auf. Links bzw. rechts positionierte Schallquellen ergeben negative bzw. positive Panning Index Werte, wobei gilt $-1 \leq \Psi \leq +1$.

Für jede Frequenz k und für jeden Zeitframe m der STFT liegt ein Panning Index Wert $\Psi[m,k]$ vor. Eine unmittelbare Verwendung dieser Werte als Merkmale für die Sprachdetektion, würde aufgrund der hohen Anzahl der Merkmale einen hohen Rechenaufwand für den Klassifizierer bedeuten. Zudem wären mehr Trainingsdaten vonnöten, da die Komplexität des Modells des Klassifizierers mit der Anzahl der Merkmale steigt.



(a) Veränderung des Schwellwertes ϕ_0

(b) Veränderung der Steilheit σ

Abbildung 4.3: Panning Index Gewichtungsfunktion $\Gamma_\Psi(\Psi)$

Um die berechneten Werte der Interkanal-Pegeldifferenz also effizient als Merkmale für den Klassifizierer nutzen zu können, werden sie über drei Frequenzbänder mit Hilfe des arithmetischen Mittels und der Standardabweichung zusammengefasst. Bei den Frequenzbändern κ_j $(j = 1, 2, 3)$ handelt es sich um 100 bis 400 Hz, 400 bis 1600 Hz sowie 1600 bis 6400 Hz. Diese Bänder wurden dem Frequenzbereich der menschlichen Sprache angepasst. Jedes dieser Bänder erstreckt sich über zwei Oktaven, das heißt die Breite der einzelnen Frequenzbänder wächst mit ansteigender Frequenz. Das arithmetische Mittel des Panning Indexes $\bar{\Psi}[m, k]$ wird für jeden STFT Zeitframe m berechnet. Summiert wird jeweils über alle Frequenzindizes k, die Teil der definierten Frequenzbänder κ_j sind.

$$\bar{\Psi}[m, \kappa_j] = \frac{1}{N_{\kappa_j}} \sum_{k \in \kappa_j} \Psi[m, k] \tag{4.4}$$

wobei es sich bei N_{κ_j} um die Anzahl der verwendeten FFT-Punkte pro Frequenzband handelt. Die Standardabweichung der Panning Index Verteilung wird für jedes Frequenzband κ_j folgendermaßen angenähert:

$$\sigma_\Psi[m, \kappa_j] = \sqrt{\sum_{k \in \kappa_j} \frac{(\Psi[m, k] - \bar{\Psi}[m, \kappa_j])^2}{N_{\kappa_j} - 1}} \tag{4.5}$$

Für jeden STFT Block m steht also jeweils für Panning Index Mittelwert und Panning Index Standardabweichung ein Wert pro Band κ_j zur Verfügung.

Tzanetakis et al. [69] benutzen den RMS-Wert (Abkürzung für Root Mean Square) des Panning Indexes $\Psi[m,k]$ über vier Frequenzbänder um den Produktionsstil von Musikaufnahmen zu klassifizieren. Durch die Verwendung des RMS Wertes kann nur die Breite des Pannings bestimmt werden, nicht aber die Richtung der gepannten Quellen. Das oberste Frequenzband erstreckt sich bis zu 22050 Hz.

Magnitudengewichtete Panning Index Merkmale

Berechnet man Stereomerkmale über bestimmte Frequenzbänder, wie es im vorherigen Abschnitt beschrieben wurde, so haben Merkmalswerte von Frequenzen kleiner Magnituden den gleichen Einfluss auf das Ergebnis des Merkmales wie Werte großer Magnituden. Der Pegel von Dialogen in Filmen ist zum Beispiel etwa 20 dB bis 40 dB lauter als der von Atmosphären [3]. Das hier entwickelte magnitudengewichtete Merkmal soll Merkmalswerten jener Frequenzen mit starker Magnitude eine größere Bedeutung zukommen lassen.

Zu diesem Zweck wird ein gewichtetes Histogramm des Merkmals erstellt. Es erfolgt eine Einteilung Wertebereichs des Merkmals in I Intervalle [70]. Fällt ein Wert nun in eines der Intervalle i, wird die Häufigkeit $h[i]$ eines herkömmlichen Histogramms um eins inkrementiert: $h[i] = h[i] + 1$. Das gewichtete Histogramm aber wird um den aus der STFT berechneten Absolutbetrag $|H[k]|$ einer Frequenz k erhöht. Es gilt:

$$\bar{h}[i] = \bar{h}[i] + |H[k]| \tag{4.6}$$

Das so entstehende Histogramm wird mit Hilfe von Momenten höherer Ordnung in seinen Eigenschaften Varianz, Schiefe und Exzess beschrieben. Grundelement der hier verwendeten Momente ist das zentrale Moment M_l der Ordnung l zum Zeitpunkt m einer Wahrscheinlichkeitsverteilung:

$$M_l[m,\kappa_j] = \mathbf{E}\{(\Psi[m,k] - \bar{\Psi}[m,k])^l\}\Big|_{k\in\kappa_j} = \sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^l \, p_\Psi[m,k] \tag{4.7}$$

Bei $p_\Psi[m,k]$ handelt es sich um die Wahrscheinlichkeit, dass die Zufallsvariable der Panning Index Verteilung den Wert $\Psi[m,k]$ annimmt. Daraus ergibt sich die Varianz

des Panning Indexes für ein Frequenzband κ_j als

$$\sigma^2_{\Psi}[m,\kappa_j] = M_2[m,\kappa_j] = \sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^2 \, p_{\Psi}[m,k] \tag{4.8}$$

Handelt es sich in einem Film um eine Szene mit vielen über die gesamte Stereobreite verteilten Schallquellen, so ist das Histogramm des Panning Indexes eher flach und erstreckt sich über den gesamten Wertebereich von $-1 \leq \Psi[m,\kappa_j] \leq +1$. Handelt es sich hingegen um eine Szene mit Sprache, deren Pegel deutlich lauter ist als die sie umgebende Atmosphäre, so wird der Varianzwert des Histogramms deutlich kleiner sein.

Die Wölbung b_2 einer Wahrscheinlichkeitsverteilung ist definiert als

$$b_2 = \frac{M_4}{M_2^2} \tag{4.9}$$

Handelt es sich um eine Verteilung mit steilem Gipfel gilt $b_2 > 0$, während für Verteilungen mit flachem Gipfel gilt $b_2 < 0$. Der Exzess g_2 einer Wahrscheinlichkeitsverteilung, steht für die Wölbung einer Verteilung in Bezug zur Wölbung einer Normalverteilung für die gilt: $b_2 = 3$. Der Exzess wird folgendermaßen berechnet:

$$g_{2,\Psi}[m,\kappa_j] = \frac{M_4}{M_2^2} - 3 = \frac{\sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^4 \, p_{\Psi}[m,k]}{\left[\sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^2 \, p_{\Psi}[m,k]\right]^2} - 3 \tag{4.10}$$

Die Schiefe g_1 beschreibt die Schrägheit einer Verteilung. Neigt sich die Panning Index Verteilung nach rechts, also zu positiven Werten hin, so besitzt sie eine negative Schiefe. Neigt sie sich nach links, so ist ihr Schiefewert positiv.

$$g_{1,\Psi}[m,\kappa_j] = \frac{M_3}{M_2^{\frac{3}{2}}} = \frac{\sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^3 \, p_{\Psi}[m,k]}{\left[\sum_{k\in\kappa_j} (\Psi[m,k] - \bar{\Psi}[m,\kappa_j])^2 \, p_{\Psi}[m,k]\right]^{\frac{3}{2}}} \tag{4.11}$$

Interkanal-Kohärenz Stereomerkmale

Die Interkanal-Kohärenz (Englisch: interchannel coherence, ICC) $\phi[m,k]$, beschreibt die Kohärenzeigenschaften zwischen linkem und rechtem Kanal eines Signals. So kann zwischen zwischen kohärenten ($\phi \to 1$) und inkohärenten ($\phi \to 0$) Signalanteilen unterschieden werden. Wie bei den Panning Index Merkmalen erfolgt auch hier eine Zusammenfassung des Frequenzbereichs in die drei Bänder κ_j ($j = 1, 2, 3$) mit 100 bis 400 Hz,

400 bis 1600 Hz sowie 1600 bis 6400 Hz. Die Standardabweichung $\sigma_\phi[m, \kappa_j]$ sowie das arithmetische Mittel $\bar{\phi}[m, \kappa_j]$ der ICC-Werte jedes Bands werden analog zu den Panning Index Merkmalen berechnet.

$$\mu_\phi[m, \kappa_j] = \frac{1}{N_{\kappa_j}} \sum_{k \in \kappa_j} \phi[m, k] \tag{4.12}$$

Der Näherungswert der Standardabweichung σ_ϕ der ICC-Verteilung über die Frequenzbänder κ_j ergibt sich dann zu:

$$\sigma_\phi[m, \kappa_j] = \sqrt{\sum_{k \in \kappa_j} \frac{(\phi[m, k] - \bar{\phi}[m, \kappa_j])^2}{N_{\kappa_j} - 1}} \tag{4.13}$$

Seiteninformation Merkmal (SIF)

Motiviert aus der MS- Stereophonie [12] wird für das Seiteninformation Merkmal (SIF) die mittlere Energie der sogenannten Seiteninformation bezüglich der mittleren Gesamtenergie eines Signals pro Frequenzband berechnet. Es werden wieder die Frequenzbänder κ_j $(j = 1, 2, 3)$ mit 100 bis 400 Hz, 400 bis 1600 Hz sowie 1600 bis 6400 Hz benutzt.

Für die MS- Stereophonie benötigt man zwei Mikrophone, die auf azimutaler Ebene um 90° zueinander versetzt ausgerichtet sind. Das Mittensignal im Zeitbereich $x_M[n]$ wird durch einen Druckempfänger mit Kugelrichtcharakteristik oder auch Druckgradientenempfänger mit Nieren- oder Achterrichtcharakteristik erzeugt. Das Mikrofon wird hierbei direkt auf die Schallquelle(n) gerichtet. Das Seitensignal $x_S[n]$ hingegen stammt immer von einem Druckgradientenempfänger mit Achterrichtcharakteristik. Durch einfache Matrizierung der beiden Signale können der linke bzw. rechte Stereokanal $x_L[n]$ und $x_R[n]$ erzeugt werden.

$$x_L[n] = \frac{1}{\sqrt{2}}(x_M[n] + x_S[n]) \tag{4.14}$$

$$x_R[n] = \frac{1}{\sqrt{2}}(x_M[n] - x_S[n]) \tag{4.15}$$

Für das Seiteninformation Merkmal geht man den umgekehrten Weg der MS- Stereophonie. Ausgehend von den existierenden Stereospuren $x_L[n]$ und $x_R[n]$ werden Seitensignal

$x_S[n]$ und Mittensignal $x_M[n]$ durch Umrechnen der Gleichungen 4.14 und 4.15 folgendermaßen erzeugt:

$$x_S[n] = \frac{1}{\sqrt{2}}(x_L[n] - x_R[n]) \tag{4.16}$$

$$x_M[n] = \frac{1}{\sqrt{2}}(x_L[n] + x_R[n]) \tag{4.17}$$

Die Linearitätseigenschaft der Fourier Transformation [26] lautet:

$$ax_L[n] + bx_R[n] \overset{F}{\leftrightarrow} aX_L[k] + bX_R[k] \tag{4.18}$$

So kann die Berechnung von Gleichung 4.16 und 4.17 im Frequenzbereich vollzogen werden:

$$X_S[m,k] = \frac{1}{\sqrt{2}}(X_L[m,k] - X_R[m,k]) \tag{4.19}$$

$$X_M[m,k] = \frac{1}{\sqrt{2}}(X_L[m,k] + X_R[m,k]) \tag{4.20}$$

Das Seiteninformation Merkmal SIF ist definiert als das Verhältnis der mittleren Energie der Seiteninformation zur mittleren Gesamtenergie aus Seiten- und Mittensignal. Es liegt für jedes Frequenzband κ_j vor.

$$SIF[m,\kappa_j] = \frac{\frac{1}{N}\sum_{k\in\kappa_j} X_S^2[m,k]}{\frac{1}{N}\sum_{k\in\kappa_j} X_M^2[m,k] + \frac{1}{N}\sum_{k\in\kappa_j} X_S^2[m,k]} \tag{4.21}$$

Ein Stereosignal mit zwei identischen Kanälen, ein Monosignal, besitzt Seiteninformationen deren mittlere Energie gegen 0 strebt, folglich strebt auch der Wert des Seiteninformation Merkmals gegen 0. Signale, deren Seiteninformationen einen hohen Anteil an der mittleren Gesamtenergie ausmachen, besitzen einen SIF Wert, der gegen 1 strebt.

4.1.3 Nachverarbeitung

Modellbasierte Klassifizierung zur Laufzeit

Das Training eines Klassifizierers wie den neuronalen Netzwerken erfolgt mit einem möglichst großen Set an Trainingsdaten, das möglichst repräsentativ für alle in den

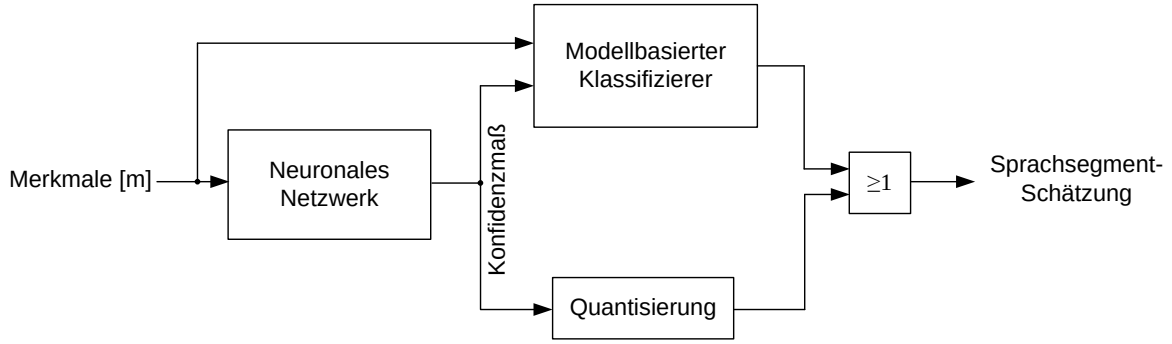


Abbildung 4.4: Einbettung des modellbasierten Klassifizierung zur Laufzeit in das Spracherkennungssystem

Filmen vorkommenden Szenarien wie verschiedenste Sprecherstimmen, Hintergrundgeräusche usw. ausgewählt wird. Die Bildung des Modells eines Klassifizierers erfolgt so, dass die Klassifizierungsgenauigkeit von Sprache für das gesamte Set mit all seinen Sprecher/innen möglichst gut ist. Wüsste man schon vor dem Training des Klassifizierers, welche Szenarien und Sprecher/innen in einem Film vorkommen, könnte man das Trainingsdatenset genau auf die Anforderungen einstellen. Die grundlegende Idee der modellbasierten Klassifizierung ist, zusätzlich zum bereits bestehenden Klassifizierer einen weiteren Klassifizierer zur Laufzeit zu trainieren, wobei die Schätzungen des ersten Klassifizierers Aufschluss darüber geben, mit welchen Daten der modellbasierte Klassifizierer trainiert werden soll.

Abbildung 4.4 zeigt die Integration dieser Nachverarbeitung in das Sprachdetektionssystem. Das neuronale Netzwerk schätzt, welcher Klasse die eingehenden Merkmale zugehören. Das Ausgangsneuron des hier verwendeten MLPs mit logistischer Aktivierungsfunktion nimmt kontinuierliche Werte zwischen 0 und 1 an. Dabei handelt es sich bei Werten, die gegen 1 streben mit großer Sicherheit um Sprache. Werte, die gegen 0 streben deuten mit großer Wahrscheinlichkeit auf Nicht-Sprache hin. Dieses Maß wird in Folge als Konfidenzmaß bezeichnet und bildet den Ausgangspunkt für die modellbasierte Klassifizierung. Die diskretisierten Schätzungen der beiden Klassifizierer werden über eine Oder- Verknüpfung miteinander verbunden. Dies bedeutet also, sobald nur einer der beiden Klassifizierer einen Frame für Sprache hält, lautet die Schätzung des

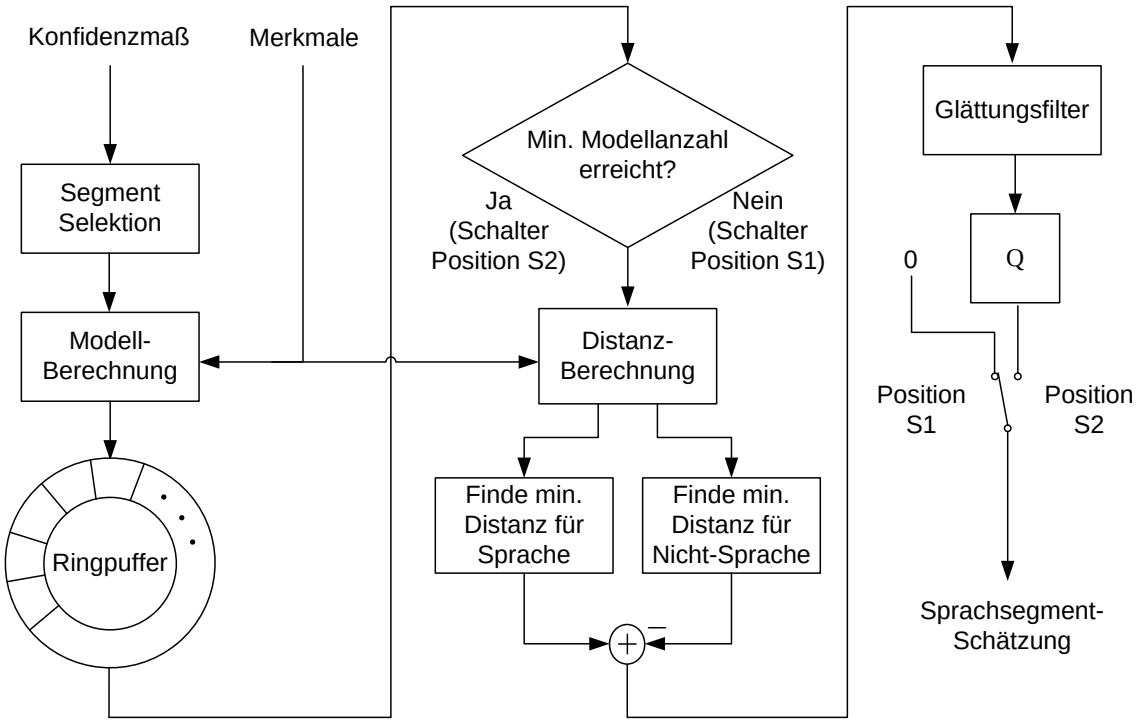


Abbildung 4.5: Blockschaltbild der modellbasierten Klassifizierung zur Laufzeit

gesamten Sprachdetektionssystems Sprache.

Abbildung 4.5 gibt einen groben Überblick über das Gesamtsystem der modellbasierten Klassifizierung zur Laufzeit.

Anhand der Konfidenzwerte erfolgt eine Selektion der für die Modellbildung geeigneten Segmente. Es werden sowohl Modelle für Sprache, als auch Modelle für Nicht- Sprache berechnet. Die Auswahl von Segmenten erfolgt nach folgenden Kriterien:

- Ein Sprachsegment zur Modellbildung beginnt, sobald der Konfidenzwert $\zeta[m]$ einen festgelegten Schwellwert ζ_a überschreitet: $\zeta[m] > \zeta_a$. Der Beginn eines Nicht-Sprache Segments liegt hingegen vor, wenn gilt: $\zeta[m] < \zeta_b$, wobei $\zeta_a > \zeta_b$.

- Sobald der Konfidenzwert des Sprachsegments den Schwellwert ζ_a unterschreitet, wird das Segment beendet. Ein Nicht- Sprache Segment endet, wenn gilt: $\zeta[m] > \zeta_b$.

- Erst wenn ein in Frage kommendes Segment die geforderte minimale zeitliche Länge überschreitet, wird aus diesem Segment letztendlich ein Modell berechnet.

Die Auswahl der Segmente spielt für die modellbasierte Klassifizierung eine äußerst wichtige Rolle, da die Robustheit dieser Nachverarbeitung von ihr abhängt. Durch die Einführung der Konfidenz- Schwellwerte ζ_a und ζ_b wird sichergestellt, dass nur Merkmale für die Modellbildung herangezogen werden, bei denen eine hohe Wahrscheinlichkeit besteht, dass es sich um Sprache bzw. nicht um Sprache handelt. Die Forderung nach einer minimalen zeitlichen Länge von Segmenten, ist motiviert durch die erwünschte Robustheit der Modelle.

Sobald ein Segmentbereich gefunden wurde, der alle Robustheitsanforderungen erfüllt, werden die Modellparameter Mittelwert μ, Varianz σ^2 und Kovarianzmatrix Σ der Merkmalwerte dieses Segments i berechnet. Die Kovarianz $\sigma_{x_1 x_2}$ von zwei Merkmalen $\mathbf{x}_1$ und $\mathbf{x}_2$ lautet [70]:

$$\sigma_{\mathbf{x}_1\mathbf{x}_2} = \mathbf{E}\{\,(\mathbf{x}_1 - \bar{\mathbf{x}}_1)\,(\mathbf{x}_2 - \bar{\mathbf{x}}_2)\,\} \tag{4.22}$$

Es gilt: $\bar{\mathbf{x}}_i = \mathbf{E}\{\mathbf{x}_i\}$ mit $\mathbf{E}$ als Erwartungswertoperator. Handelt es sich um einen D-dimensionalen Merkmalsraum, so stellt die Kovarianzmatrix Σ die Kovarianzen aller möglichen Paare von Merkmaldimensionen dar.

$$\Sigma = \begin{bmatrix} \mathbf{E}\{(\mathbf{x}_1 - \bar{\mathbf{x}}_1)(\mathbf{x}_1 - \bar{\mathbf{x}}_1)\} & \mathbf{E}\{(\mathbf{x}_1 - \bar{\mathbf{x}}_1)(\mathbf{x}_2 - \bar{\mathbf{x}}_2)\} & \cdots & \mathbf{E}\{(\mathbf{x}_1 - \bar{\mathbf{x}}_1)(\mathbf{x}_D - \bar{\mathbf{x}}_D)\} \\ \mathbf{E}\{(\mathbf{x}_2 - \bar{\mathbf{x}}_2)(\mathbf{x}_1 - \bar{\mathbf{x}}_1)\} & \mathbf{E}\{(\mathbf{x}_2 - \bar{\mathbf{x}}_2)(\mathbf{x}_2 - \bar{\mathbf{x}}_2)\} & \cdots & \mathbf{E}\{(\mathbf{x}_2 - \bar{\mathbf{x}}_2)(\mathbf{x}_D - \bar{\mathbf{x}}_D)\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}\{(\mathbf{x}_D - \bar{\mathbf{x}}_D)(\mathbf{x}_1 - \bar{\mathbf{x}}_1)\} & \mathbf{E}\{(\mathbf{x}_D - \bar{\mathbf{x}}_D)(\mathbf{x}_2 - \bar{\mathbf{x}}_2)\} & \cdots & \mathbf{E}\{(\mathbf{x}_D - \bar{\mathbf{x}}_D)(\mathbf{x}_D - \bar{\mathbf{x}}_D)\} \end{bmatrix} \tag{4.23}$$

Liegen Modellparameter vor, werden sie in einen Ringpuffer geschrieben. Dieser lässt nur die Speicherung einer begrenzten Anzahl von Modellen zu. Während eines langen Films käme es ohne den Ringpuffer zur Speicherung einer großen Anzahl von Modellen, deren Berechnung und darauf folgender Auswertung unter Umständen ein Problem für die Echtzeitanforderung werden könnte.

Ist eine minimale Anzahl an Sprachmodellen sowie Nicht- Sprachmodellen erreicht, so wird die modellbasierte Segmentierung aktiviert und hat einen Einfluss auf die Schätzung der Sprachsegmente. Auch hier wieder ist die Forderung nach einer minimalen Anzahl von Modellen motiviert durch die Robustheit der Differenzberechung zu den Modellen.

Die Distanz eines Merkmalvektors $\mathbf{x}[m]$ zum Zeitpunkt m zu den Modellen von Sprache und Nicht-Sprache soll Aufschluss darüber geben welcher Klasse er angehört. Die Überlegung ist hierbei, dass ein Merkmalvektor zu jener Klasse gehört zu dessen Modell er den geringeren Abstand besitzt.

Für die Berechnung des Abstandes zwischen dem Mittelwert einer mulitvariaten Verteilung μ und einem Vektor $\mathbf{x}$ kann mit verschiedenen Distanzmaßen gearbeitet werden. Bei einem der verwendeten Distanzmaße handelt es sich um die Minkowski Distanz, auch P- Norm- Distanz genannt.

$$D(\mathbf{x}, \mu) = \left[\sum_{i=1}^{N} |x_i - \mu_i|^p \right]^{\frac{1}{p}} \tag{4.24}$$

Durch die Wahl unterschiedlicher Parameterwerte p ändert sich die Gewichtung der Werte für die Distanzberechnung. Für $p = 2$ ist die Minkowski Distanz identisch der Euklidischen Distanz. Ein Distanzmaß, welches nicht nur den Abstand zum Mittelwertsvektor einer Verteilung berechnet, sondern zusätzlich die Kovarianzmatrix $\mathbf{\Sigma}$ einer Verteilung mit einbezieht, ist die Mahalanobis Distanz. So wird der Abstand eines Vektors zu einer Verteilung in Abhängigkeit zur Kovarianz gewichtet.

$$D_{\lambda_i}(\mathbf{x}) = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)} \tag{4.25}$$

Die Vorteile der Mahanalobis Distanz kann man anhand eines Beispiels für den eindimensionalen Fall verdeutlichen. Es soll der Abstand eines Punktes x zu einer Normalverteilung mit kleiner Varianz und einer weiteren mit großer Varianz ermittelt werden. Beide Verteilungen haben den selben Mittelwert, daher ergibt die Euklidische Distanz den gleichen Distanzwert zu beiden Verteilungen. In die Mahalanobis Distanz jedoch fließen zusätzlich die Kovarianzeigenschaften einer Verteilung ein. Aus Gl. (4.25) folgt, dass der Abstand zu jener Verteilung größer ist, die die kleinere Varianz besitzt. Nachdem die Distanzen eines Merkmalvektors zu allen gespeicherten Sprach- und Nicht- Sprachmodellen berechnet wurde, werden gemäß Gl.(4.26) die Distanzen zu den nächstgelegenen Verteilungen von Nicht-Sprache und Sprache voneinander subtrahiert.

$$\gamma(\mathbf{x}) = \min_i \{D_{\bar{\lambda_i}}(\mathbf{x})\} - \min_i \{D_{\lambda_i}(\mathbf{x})\} \tag{4.26}$$

Ist der Wert γ positiv, so ist der Abstand eines Merkmalvektors $\mathbf{x}$ zu einem Nicht-Sprache Modell größer als zu einem Sprache Modell. Es würde in diesem Fall von der modellbasierten Klassifizierung geschätzt werden, dass es sich bei diesem Vektor um Sprache handelt. Negative Werte von γ deuten auf Nicht- Sprache hin. Das Ergebnis $\gamma(\mathbf{x})$ wird zeitlich mittels Tiefpassfilterung geglättet und darauf durch einen Quantisierer bearbeitet. Das quantisierte Signal $\gamma_Q(\mathbf{x})$ mit Schwellwert γ_{th} ergibt sich aus folgendem Zusammenhang:

$$\gamma_Q(\mathbf{x}) = \begin{cases} 1 & falls \quad \gamma(\mathbf{x}) > \gamma_{th} \\ 0 & sonst \end{cases} \tag{4.27}$$

Hüllkurvensegmentierung mittels adaptiver Hintergrund-Pegelberechnung

Das bestehende Upmixsystem soll durch Überblendung zwischen zwei Klangeinstellungen verbessert werden. Diese Klangeinstellungen sind jeweils angepasst an Sprache und an Inhalte, die keine Sprache enthalten wie z.B. Musik, Atmosphären und Effekte. Der Zeitpunkt, wann diese Überblendung erfolgt, hängt vom Sprachdetektionsergebnis ab. Die zeitliche Genauigkeit der Segmentgrenzen ist daher für den Upmixer, dessen Überblendungen auf den Segmentinformationen basieren, von großer Bedeutung.

Bei der Hüllkurvensegmentierung handelt es sich um eine Nachverarbeitung der Sprachsegmentschätzungen des Klassifizierers. Es soll eine zeitliche Verbesserung des On- und Offsets, also des Beginns und des Endes eines Sprachsegments erreicht werden. Denn selbst wenn ein Klassifizierer ein Sprachsegment innerhalb eines Signals erkennt, kann die zeitliche Genauigkeit der Segmentgrenzen noch verbessert werden.

Liegen Sprache und Atmosphären zur gleichen Zeit vor, kann man davon ausgehen, dass der Sprachpegel in Filmen in der Regel etwa 20 dB bis 40 dB höher als der Pegel von Atmosphären ist [3]. Sprache, insofern sie der Handlung beiträgt, soll verständlich sein und hebt sich daher hinsichtlich des Pegels vom Hintergrund ab. Die Hüllkurvensegmentierung untersucht, ob der aktuelle Pegel um einen gewissen Schwellwert über dem Hintergrundpegel befindet. Folgende Gleichung muss also erfüllt werden:

$$L[m] > L_{Hintergrund}[m] + L_{Schwelle}[m] \tag{4.28}$$

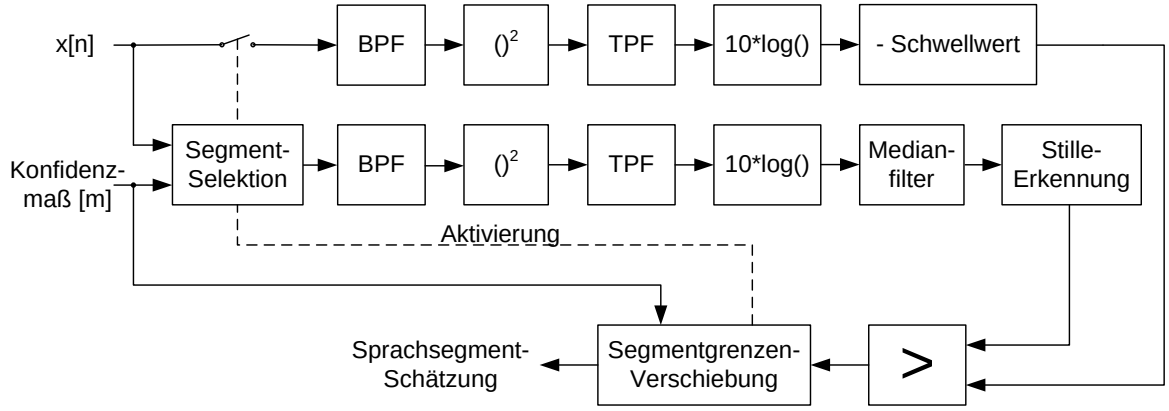


Abbildung 4.6: Blockschaltbild der Hüllkurvensegmentierung mittels adaptiver Pegelberechnung

Es findet keine Unterscheidung zwischen Sprache oder anderen Klassen macht. Es geht zunächst ausschließlich um die Erkennung von Ereignissen, die einen erhöhten Pegel aufweisen.

Nimmt man ein statisches Signal an, das in seiner Lautstärke während eines ganzen Films konstant ist, kann man einen statischen Wert für den Hintergrundpegel wählen. Sobald sich aber der Hintergrundpegel über den Film hin verändert, muss er adaptiv berechnet werden. Die Berechnung erfolgt mit Hilfe von Hinweisen des Klassifizierers in Form des Konfidenzmaßes.

Abbildung 4.6 gibt einen Überblick über die Funktionsweise der Hüllkurvensegmentierung. Die Berechnung des Hintergrundpegels eines Films erfolgt adaptiv mit Hilfe des Konfidenzmaßes des Klassifizierers. Der Hintergrundpegel eines Signals wird lediglich für Abschnitte berechnet, deren Konfidenzmaß sich für eine geforderte Mindestdauer unter einem Schwellwert befindet. So soll sicher gestellt werden, dass keine Sprachanteile in die Hintergrundpegelberechnung mit einfließen.

Die Hüllkurvensegmentiertung beginnt erst in die Schätzung der Sprachsegmentgrenzen einzugreifen, wenn ein Hintergrundpegel berechnet werden konnte. Falls es sich um einen sehr langen Nicht- Sprache Bereich handelt, wird ein Segment automatisch nach einer wählbaren Zeit beendet und der Pegel berechnet. Dies hat den Vorteil, dass die

Segmentierung früher in die Schätzung eingreifen kann.

Zur Berechnung des Pegels wird nur ein Frequenzbereich von 200-4000 Hz betrachtet. Hierfür wird das Signal mit einem elliptischen Bandpass zweiter Ordnung gefiltert, der alle Frequenzen ober- und unterhalb diese Frequenzbereichs dämpft. Danach werden die Werte im Zeitbereich quadriert und somit die Energie der Samples eines Signals berechnet. Die errechneten Werte werden zur Glättung der entstehenden Kurve mittels Butterworth Tiefpassfilter zweiter Ordnung, der eine Cutoff- Frequenz von 10 Hz besitzt, bearbeitet. Um den Pegel eines Nicht- Sprache Segments zu errechnen folgt zum Schluss noch eine Logarithmierung und Mittelung über das ausgewählte Segment mit einem Medianfilter. Der Medianfilter bekommt den Vorzug vor einem Filter, der einen arithmetischen Mittelwert berechnet, weil er kurze Extremwerte ausblenden kann und damit die Pegelberechnung robust gegenüber kurzzeitigen transienten Signalen ist. Die nachfolgende Erkennung von Stille verhindert, dass der Hintergrundpegel zu niedrig angesetzt wird und somit der Pegel aller nachfolgenden Samples die Gleichung 4.28 leicht erfüllen kann. Bei der Detektion von Stille im Signal, wird der Pegel aller Samples, die weniger als -60 dB aufweisen, auf -60 dB gesetzt.

Sobald ein Hintergrundpegel eines Abschnitts vorliegt, wird der Pegel des Signals $x[n]$ für jeden Abtastzeitpunkt n berechnet und die Verschiebung der Segmentgrenzen aktiviert. In Abbildung 4.7 sieht man die Funktionsweise dieser Verarbeitung. Die rote, strichliert dargestellte Funktion stellt das kontinuierliche Konfidenzmaß des Klassifizierers dar. Überschreitet das Konfidenzmaß einen Konfidenzschwellwert, wird ermittelt, ob die Pegelgleichung 4.28 vor bzw. nach diesem Bereich erfüllt ist. Sollte dies der Fall sein, wird das durch das Konfidenzmaß angezeigte Sprachsegment zeitlich so weit ausgedehnt, bis die Pegelgleichung nicht mehr erfüllt oder eine maximale Verschiebungszeit erreicht wird. Zeigt ein Konfidenzmaß nicht an, dass es sich bei einem Segment um Sprache handeln könnte, so findet auch keine Verschiebung der Segmentgrenzen statt.

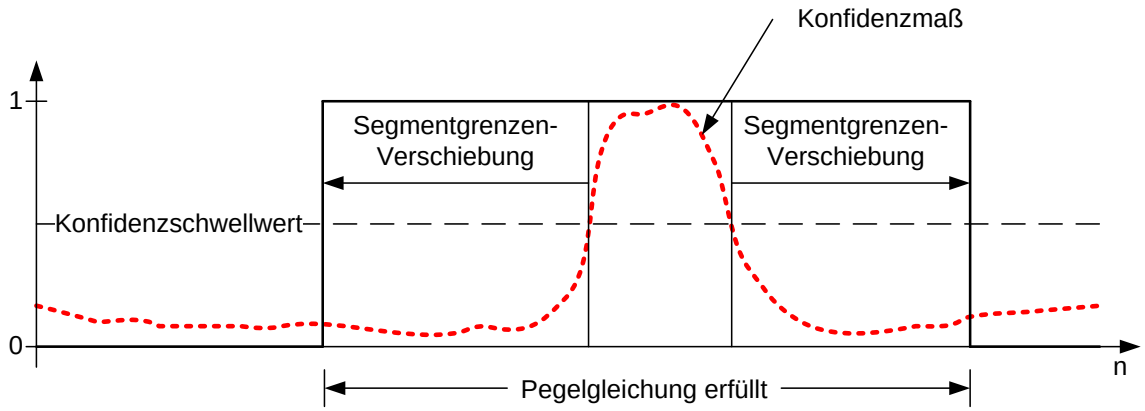


Abbildung 4.7: Verschiebung der Segmentgrenzen durch die Hüllkurvensegmentierung

4.2 Upmixer

Im Rahmen der Arbeit wurde der Ansatz verfolgt, basierend auf den Ergebnissen einer Sprachdetektion die Klangparameter des Upmixers zu steuern. Es sollte vor Allem die Qualität des klanglichen Ausgangszustands des Upmixers, hinsichtlich der Wiedergabe von Sprache in TV- und Filmton verbessert werden. Ziel war, eine klanglich unverfälschte Wiedergabe von Sprache aus dem Centerkanal zu erreichen.

4.2.1 Bestimmung der Klangeinstellung für Sprache

Der klangliche Ausgangszustand des Upmixers, soll der Wiedergabe von Musik, Atmosphären und allen Signalanteilen, bei denen es sich nicht um Sprache handelt, dienen. Diese Klangeinstellung wird als Music Mode bezeichnet.

Es soll eine Klangeinstellung gefunden werden, die sich besonders zur Wiedergabe von Sprache eignet. Durch welche Eigenschaften zeichnet sich jedoch eine geeignete Klangeinstellung aus? Es wurden informelle Hörtests durch den Autor durchgeführt, um die Ursachen der Schwächen des Upmixers bezüglich der Wiedergabe von Sprache zu ermitteln. Es stellte sich dabei heraus, dass die Qualität der Sprachwiedergabe von Parametern, die die Sprachwiedergabe aus den äußeren Lautsprechern unterdrückten, günstig beeinflusst wurde. Wird Sprache über alle drei Frontlautsprecher wiedergegeben, kommt

es im Vergleich zur 5.1 Studiomischung zu einer unpräzisen räumlichen Darstellung sowie der klanglichen Verfärbung von Sprache. Eine Ursache der breiteren räumlichen Darstellung von Sprache liegt darin, dass Sprache durch die Aktivität der linken und rechten Frontlautsprecher als virtuelle Schallquelle in der Mitte dargestellt wird. Virtuelle Schallquellen haben die Eigenschaft, weniger scharf als reale Schallquellen lokalisiert zu sein. Darüber hinaus werden häufig Anteile des Hörereignisses der virtuellen Schallquelle aus unterschiedlichen Richtungen lokalisiert [6].

Der Upmixer arbeitet mit einer zeitlichen Verzögerung der Signale der äußeren Lautsprecher, bezogen auf den Centerkanal. Aus der Raumakustik ist bekannt: Trifft ein hoher Anteil der frühen Reflexionen (bis zu 80 ms) einer Schallquelle aus seitlichen Richtungen beim Hörer ein, ergibt sich eine breitere räumliche Wahrnehmung einer Schallquelle [71]. Für den Music Mode ist diese räumliche Darstellung von Schallquellen erwünscht. Doch ergeben sich dadurch auch klangliche Verfärbungen, aufgrund des entstehenden Kammfilters [72].

Durch eine Klangeinstellung, die möglichst wenig Sprachanteile auf den äußeren Frontlautsprechern beinhaltet, kann also einerseits die räumliche Abbildung, sowie die klangliche Wiedergabe von Sprache verbessert werden. Darüber hinaus kann Sprache auch von nicht idealen Hörpositionen deutlich aus Richtung des Centerlautsprecher lokalisiert werden.

4.2.2 Center Integration

Es wurden durch den Autor informelle Hörtests mit mehreren Klangparametern, die die oben genannten Anforderungen erfüllen, durchgeführt. Als wirksamer Parameter kristallisierte sich der Parameter *Center Integration* mit einem Wertebereich von 0 bis 1 heraus. Das Centersignal wird mit dem Parameter multipliziert. Gleichzeitig werden Anteile des Centerkanals so auf die äußeren Frontkanäle L und R addiert, dass die Gesamtlautstärke der drei Frontkanäle gleich bleibt. Über den Parameter *Center Integration* kann also festgelegt werden, wie stark der Centerlautsprecher an der Wiedergabe beteiligt werden soll. Der Wert der *Center Integration* für den Music Mode beträgt 0.5, d.h. Teile des

Centersignals werden zu den äußeren Frontlautsprechern hinzugefügt.

Für die Erstellung des Speech Mode, musste die *Center Integration* so eingestellt werden, dass Übergänge zwischen den beiden Klangeinstellungen Movie- und Speech-Mode so wenig wie möglich hörbar sind. Gleichzeitig sollte jedoch eine Verbesserung der Wiedergabe von Sprache erreicht werden. Für die Hörtests in Kapitel 5.2 wurde ein Wert von 0.7 gewählt.

4.2.3 Berechnung des Steuersignals des Upmixers

Die Sprachdetektion liefert ein diskretes Ergebnis für jedes Sample eines Audiosignals, das für die An- oder Abwesenheit von Sprache innerhalb eines Audiosignals stehen. Benutzt man dieses Klassifikationsergebnis unmittelbar zur Steuerung der Klangeinstellungen des Upmixers, so erhöht sich die Gefahr von hörbaren Umschaltvorgängen. Um dies zu vermeiden wurde für die vorliegende Arbeit eine lineare Überblendung zwischen den diskreten Werten der Klassen Sprache und Nicht- Sprache realisiert.

Wichtige Parameter sind hierbei die Ein- und Ausblendzeit, innerhalb der eine Überblendung zwischen zwei Zuständen vollzogen wird. Zwei Varianten mit unterschiedlichem Bedarf an Lookahead (Deutsch: Vorgriffszeit) wurden implementiert. Die Verwendung eines Lookaheads erlaubt es, zum Zeitpunkt n der Verarbeitung nicht nur das aktuelle Sample $x[n]$ und zeitlich in der Vergangenheit liegende Samples, sondern auch Samples $x[n+L]$ die sich um eine endliche Lookahead-Zeit L in der Zukunft befinden, zu erfassen und auszuwerten.

Abbildung 4.8 zeigt die Auswirkungen von Lookahead und Echtzeitvariante auf die Überblendung. In der oberen Darstellung sieht man die detektierten Sprachsegmente über der Zeit. Die mittlere bzw. untere Darstellung zeigt die resultierenden Lookahead- sowie Echtzeit Varianten der Überblendung.

Die Lookahead Variante kann sich auf ein auftretendes Sprachsegment in der Zukunft einstellen und beginnt vor dem Beginn des Sprachsegments mit der Überblendung. Die Echtzeitvariante hingegen beginnt die Überblendung erst zu Beginns eines Sprachsegments. Die beiden Varianten sind, was das Ausblenden eines Segments angeht, identisch.

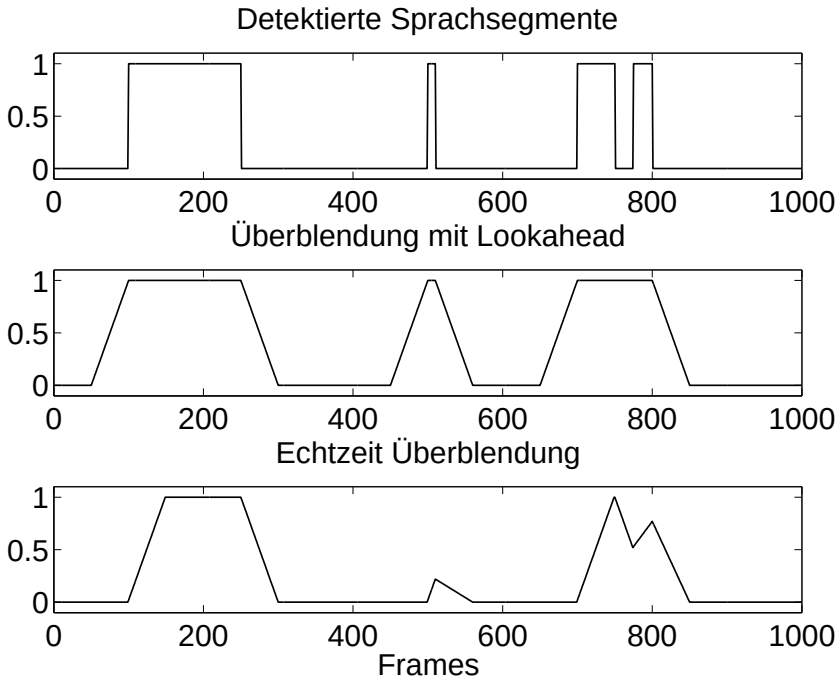


Abbildung 4.8: Überblendung zwischen Ergebnissen der Sprachdetektion

Sprachsegmente, deren zeitliche Ausdehnung kleiner als die Einblendzeit ist, werden durch die Überblendung in Echtzeit praktisch im Music Mode wiedergegeben, wie in Abbildung 4.8 zum Zeitpunkt von 500 Frames gezeigt. Das Sprachsegment weist in diesem Bereich eine Breite von 10 Frames auf, während die Einblendzeit bei 50 Frames liegt. Die Überblendung in Echtzeit blendet hier über eine Dauer von 10 Frames ein, bis die Information vorliegt, dass das Segment bereits beendet ist. Darauf erfolgt die Ausblendung zum Nicht- Sprache Zustand hin. Diese Eigenschaft kann man sich im Hinblick auf die Unterdrückung von Fehlklassifikationen deren Dauer sich im entsprechenden Bereich liegt zu Nutze machen.

Die Überblendung mit Lookahead hingegen verfügt nicht über diese Unterdrückungseigenschaft. Zum Zeitpunkt des angezeigten Sprachsegments ist die Überblendung vollständig ausgeführt.

Zeitlich kurz aufeinander folgende Segmente wie in Abbildung 4.8 im Zeitbereich von 700 bis 800 Frames gezeigt, werden durch die Überblendung mit Lookahead vereinigt,

da sich die Werte von Ein- und Ausblendung additiv überlagern. Die Echtzeitvariante wurde für diesen Fall so konzipiert, dass sie solange einblendet oder ausblendet, bis sie durch in der Folge auftretende Sprachsegmentgrenzen dazu gebracht wird ihre Richtung zu ändern. Dadurch wird ebenfalls bis zu einem gewissen Grad eine Vereinigung der Segmente erreicht.

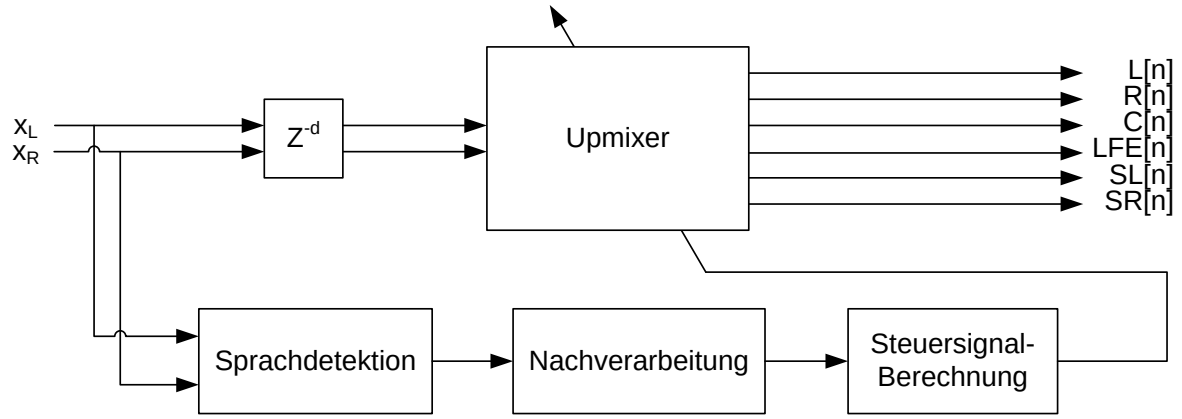


Abbildung 4.9: Überblick über die Arbeitsweise des Gesamtsystems

4.3 Gesamtsystem

Die in den Kapiteln 4.1 und 4.2 beschriebenen Algorithmen werden, wie in Abbildung 4.9 visualisiert, zu einen Gesamtsystem zusammen gefügt. Eine Sprachdetektion erkennt die Anwesenheit von Sprache innerhalb einer Stereotonspur. Die ermittelten Sprachsegmente werden darauf folgend durch eine Hüllkurvensegmentierung bearbeitet. Einerseits unterdrückt diese Nachverarbeitung sehr kurz auftretende Sprachsegmente, andererseits verbreitert sie die nicht unterdrückten Segmentgrenzen und orientiert sich dabei an der Hüllkurve des vorliegenden Tonmaterials (siehe Abschnitt 4.1.3).

Aus den nachverarbeiteten Ergebnissen der Sprachdetektion erfolgt die Berechnung eines Steuersignals wie in Abschnitt 4.2.3 beschrieben. Hierfür wird eine Steuerfunktion, die zwischen den diskreten Zuständen Sprache und Nicht- Sprache überblendet, erzeugt. Es können verschiedene Überblendparameter wie die Überblendzeit und Überblendcharakteristik eingestellt werden. Weiterhin kann bestimmt werden, ob die Überblendungen in Echtzeit oder mit einer bestimmten Lookahead-Zeit ausgeführt werden sollen. Mit der Verwendung eines Lookaheads ergibt sich automatisch eine Latenz des Gesamtsystems: Das Eingangssignal liegt erst nach einer Verzögerungszeit am Ausgang des Systems vor.

Die Upmix- Einheit besitzt eine Schnittstelle durch die ein sogenannter Metaparame-

ter, dessen Endpunkte jeweils eine für Sprache und Nicht-Sprache angepasste Klangeinstellung repräsentieren, für jeden Zeitpunkt m mit kontinuierlichen Werten zwischen 0 und 1 angesteuert werden kann. Diese Metaparameter- Steuerung kontrolliert dabei alle ausgewählten Parameter, die sich durch eine Überblendung zwischen Sprach- und Nicht-Sprach- Klangeinstellung ändern.

Durch die Wahl bestimmter Einstellungen in den Modulen Segment- Nachverarbeitung und Berechnung des Steuersignals kann es dazu kommen, dass das Gesamtsystem nicht mehr echtzeitfähig ist. Der Upmixer muss deshalb eine entsprechende Zeitverzögerung z^{-d} gegenüber dem Steuerungszweig besitzen.

4.3.1 Latenzabschätzung des Gesamtsystems

Es kann von Vorteil sein, Module des in Abbildung 4.9 dargestellten Gesamtsystems mit einem begrenzten Lookahead L und der damit einhergehenden Latenz zu betreiben. Verwendet man Module auf diese Weise muss der Upmixer das Audiosignal mit einer Zeitverzögerung von d Samples verarbeiten bis das Steuersignal berechnet wurde.

Verschiedene Einstellungen von Segmentnachverarbeitung und der abschließenden Berechnung des Steuersignals können Latenz verursachen. Die Entstehung dieser Zeiten zeigt Abbildung 4.10. Als Ausgangspunkt für die Nachverarbeitung dient das Konfidenzmaß des Klassifizierers. Überschreitet das Konfidenzmaß einen wählbaren Schwellwert, so muss es sich mindestens für die Dauer einer minimalen Konfidenzzeit über dem Schwellwert befinden. Ansonsten wird es durch die Nachverarbeitung ignoriert und kein Sprachsegment angezeigt. Das System muss demnach für die Dauer der minimalen Konfidenzzeit abwarten, um wirklich sicher sein zu können, dass es sich um ein Segment handelt, dass den Anforderungen an die Mindestdauer genügt. Sobald es sich um ein gemäß den Forderungen zulässiges Segment handelt, kann die Hüllkurvennachverarbeitung damit beginnen das Segment auf bereits vergangene Werte auszudehnen. Diese Verarbeitung wird im Bild durch die Pre-Segmentzeit veranschaulicht.

Ist das Segment durch Sprachdetektion und Nachverarbeitung bestimmt, kommt es durch die Einblendzeit, innerhalb der ein Übergang zwischen einem Nicht- Sprache Zu-

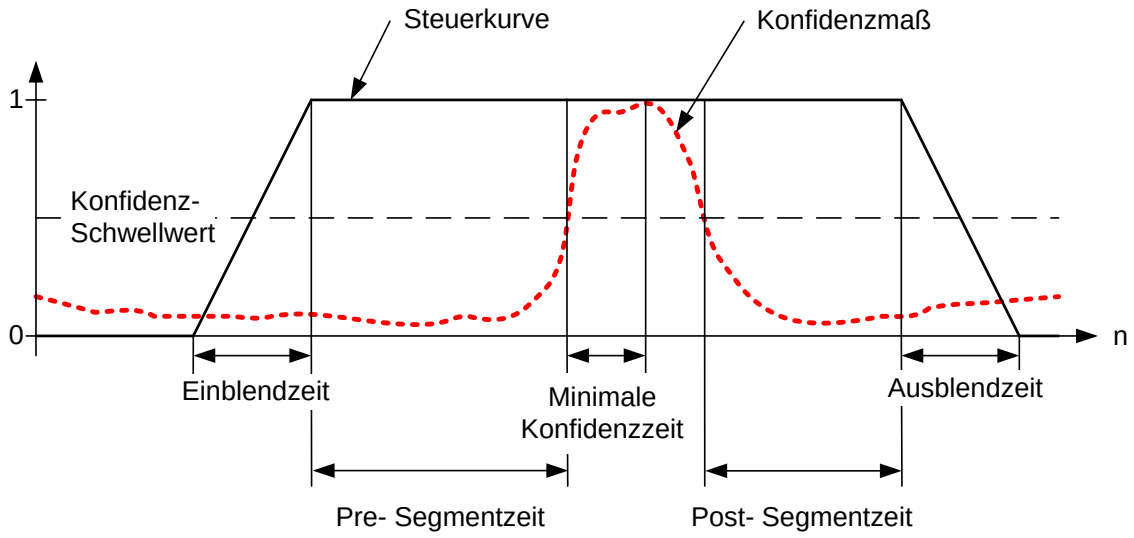


Abbildung 4.10: Einfluss der Zeitparameter auf die Entstehung der Steuerkurve

stand und einem Sprache Zustand stattfindet, zu einer weiteren Erhöhung der erforderlichen Zeitverzögerung hinsichtlich des Upmixers. Wie in Abschnitt 4.2.3 beschrieben, existieren für die Berechnung des Steuersignals des Upmixers zwei verschiedene Modi. Die Einblendzeit muss nur im Falle der Benutzung des Lookahead Modus zur Gesamtlatenz hinzugerechnet werden. Die Ausblendzeit spielt keine Rolle für die Latenz, da zu einem Zeitpunkt n feststeht, dass ein Sprachsegment beendet ist und darauf hin nur noch ausgeblendet werden muss. Es ist also keine vorausschauende Überblendung wie für den Beginn eines Sprachsegments erforderlich.

Durch die Verwendung von Sigma Merkmalen, wie in Abschnitt 3.1.3 beschrieben, kann die Genauigkeit der Sprachdetektion verbessert werden. Sigma Merkmale repräsentieren einen Block von zeitlich aufeinander folgenden Merkmalen mit Hilfe eines Mittelwerts, der durch IIR-Tiefpassfilterung angenähert wird. Die Filterung stellt eine Ursache für die Latenz der Sprachdetektion dar. Abgesehen von der Zeitverzögerung zwischen Steuerungszweig und Upmixer, muss die Latenz des Upmixers selbst zur Berechnung der Gesamtlatenz des Systems addiert werden. Das Bildmaterial muss in Folge um den Wert der Gesamtlatenz in Bezug auf das Audiosignal verzögert werden, um eine synchrone Wiedergabe von Bild und Ton zu erreichen.

Quantitative Angaben zur Latenzabschätzung befinden sich in Abschnitt 5.2.1 im Zuge der Vorstellung der im Rahmen des abschließenden Hörtests verwendeten Testsignale.

5 Ergebnisse und Diskussion

5.1 Sprachdetektion

Dieses Kapitel zeigt die Ergebnisse der Evaluation der in Kapitel 4.1 gemachten algorithmischen Erweiterungen im Bereich der Sprachdetektion. Die in diesem Kapitel verwendeten Abkürzungen der Stereomerkmale werden in Tabelle 5.1 erläutert.

5.1.1 Berechnung der Ergebnisse

Trainings- und Testdatenset

Die für alle Experimente verwendeten Test- bzw. Trainingsdaten (siehe Mustererkennung in Kapitel 2.1) bestehen aus etwa 400 verschiedenen Dateien, mit einer Dateilänge von circa 60 Sekunden. Die Daten stammen zu jeweils 20% aus den Bereichen Film, Fernsehen, reine Sprachdaten, Geräusche und instrumentale Musik. Die Filmdaten wurden aus über 30 Filmen mit 5 unterschiedlichen Sprachen (Deutsch, Englisch, Spanisch, Italienisch, Japanisch) entnommen. Zu den Dateien existieren manuell erstellte Kennzeichnungen, aus denen hervor geht, zu welchen Zeitpunkten Sprache vorliegt. Diese Kennzeichnungen dienen als Referenzen, die zur Evaluierung der Sprachdetektion verwendet werden.

Klassifikation

Im Rahmen der Arbeit wurde ein MLP der Netlab Toolbox [25] verwendet. Dieses besteht aus einer verdeckten Schicht, die 10 Neuronen mit logistischen Aktivierungsfunktionen

Merkmal	**Beschreibung**
ICCm	Mittelwert über Interkanal-Kohärenz in drei Bändern.
ICCstd	Standardabweichung über Interkanal-Kohärenz in drei Bändern.
PI1m	Mittelwert über Panning Index nach Avendano in drei Bändern.
PI1std	Standardabweichung über Panning Index nach Avendano in drei Bändern.
PI2m	Mittelwert über Panning Index Centroid in drei Bändern.
PI2std	Standardabweichung über Panning Index Centroid in drei Bändern.
AWPI1var	Varianz des Histogramms des amplitudengewichteten Panning Indexes nach Avendano.
AWPI1kur	Wölbung des Histogramms des amplitudengewichteten Panning Indexes nach Avendano.
AWPI1skw	Schiefe des Histogramms des amplitudengewichteten Panning Indexes nach Avendano.
AWPI1spd	Spread des Histogramms des amplitudengewichteten Panning Indexes nach Avendano.
AWPI2var	Varianz des Histogramms des amplitudengewichteten Panning Index Centroid.
AWPI2kur	Wölbung des Histogramms des amplitudengewichteten Panning Index Centroid.
AWPI2skw	Schiefe des Histogramms des amplitudengewichteten Panning Index Centroid.
AWPI2spd	Spread des Histogramms des amplitudengewichteten Panning Index Centroid.

Tabelle 5.1: Abkürzungen der Stereomerkmale und ihre Bedeutung.

beinhaltet. Das MLP wurde mit Hilfe des *Scaled Conjugate Gradients* Algorithmus [73] der Netlab Toolbox trainiert. Die maximale Anzahl der Iterationen wurde hierbei auf 100 begrenzt. Die Parameter des MLPs wurden in Vortests bestimmt.

Kreuzvalidierung

Für die k-fache Kreuzvalidierung wird eine Datenmenge in k möglichst gleich große Teilmengen aufgeteilt. Es werden k Test- und Trainingsdurchgänge durchgeführt, wobei jeweils eine Teilmenge als Testmenge und die restlichen $k-1$ Teilmengen als Trainingsmenge verwendet werden. Dabei wird eine Teilmenge während der gesamten Kreuzvalidierung nur einmal getestet [74]. Die 10-fache stratifizierte Kreuzvalidierung hat sich bei der Validierung von Lernmodellen [42] und der Schätzung der Genauigkeit von Klassifizierern bewährt [75]. Bei der Stratifikation werden die Daten bei der Bildung der Teilmengen so angeordnet, dass jede Teilmenge das gesamte Datenset repräsentiert. Dadurch wird die Varianz zwischen den Validierungsdurchgängen minimiert. Als Fehlermaß wurde das Verhältnis der falsch klassifizierten Frames, bezogen auf die Gesamtanzahl der Frames berechnet.

Boxplots

Für die Darstellung der Fehler wird, soweit nicht anders angegeben, der Boxplot verwendet. Die Box wird durch das erste Quartil q_1 sowie das dritten Quartil q_3 der Fehlerverteilung der Items begrenzt. Der obere Whisker umfasst maximal einen Bereich von q_3 bis $q_3+1.5(q_3-q_1)$, der untere Whisker einen Bereich von $q_1-1.5(q_3-q_1)$ bis q_1. Die Einkerbungen in der Box, die sogenannten Notches, visualisieren das 95%-Konfidenzintervall um den Medianwert. Überlappen sich die Konfidenzintervalle zweier Verteilungen nicht, so spricht man von einer statistisch signifikanten Veränderung des Medianwertes. Die approximative Berechnung der Standardabweichung basiert auf einer Normalverteilung und wird berechnet als [76]:

$$s = \frac{1.25(q_3 - q_1)}{1.35}\sqrt{N} \tag{5.1}$$

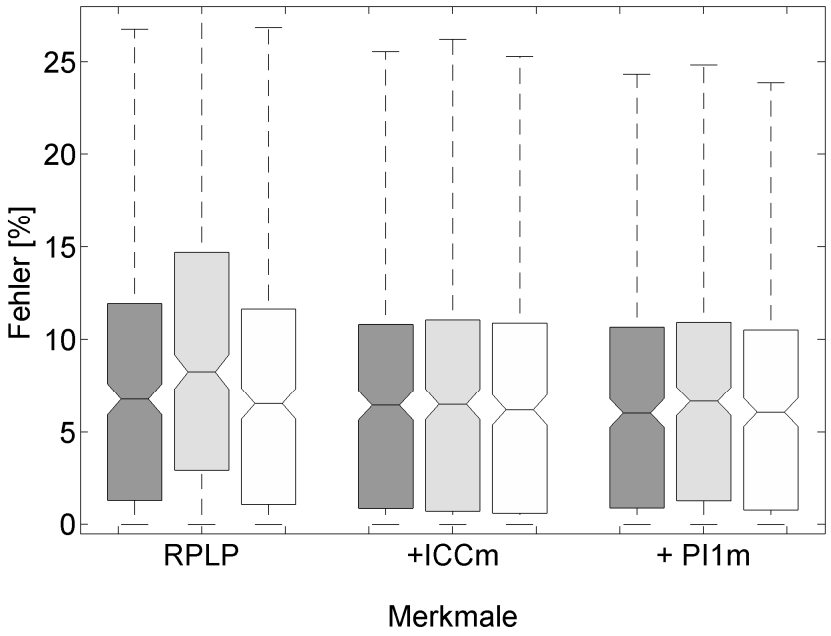


Abbildung 5.1: Fehler ohne Vorverarbeitung (dunkelgrau) sowie mit ICC-Gewichtung (hellgrau) und PI-Gewichtung (weiß) als Vorverarbeitung.

mit N als Anzahl der Items. Das Konfidenzintervall um den Wert M des Medians lautet dann:

$$M \pm 1.7s \tag{5.2}$$

Die Konstante 1.7 wurde laut [76] empirisch ermittelt. Auf eine Darstellung der Extremwerte der Fehlerverteilung durch den Boxplot wurde aus Gründen der graphischen Darstellung verzichtet, da sich die Veränderung der Medianwerte in kleinen Bereichen abspielt.

5.1.2 Spektrale Gewichtung

Die als Vorverarbeitung realisierte spektrale Gewichtung wurde entwickelt, um für die Sprachdetektion irrelevante Signalanteile mit Hilfe von spektraler Gewichtung zu unterdrücken. In Abbildung 5.1 sieht man die Boxplots, aufgetragen über den verwende-

ten Merkmalen. Bei den Merkmalen handelt es sich im ersten Fall um RASTA-PLP[1], im zweiten und dritten Fall um RASTA-PLP mit den gemittelten Stereomerkmalen Interkanal-Kohärenz (ICCm) bzw. Panning Index nach Avendano (PI1m). Wie man dem Diagramm entnehmen kann, wird durch die Vorverarbeitung mit RASTA-PLP Merkmalen und PI1-Gewichtung nur eine minimale Verbesserung des Medianwertes von 6.8 % auf 6.5 % erreicht. Diese beträgt etwa 0.2 % bei der zusätzlichen Verwendung des ICCm-Merkmals. Bei der gemeinsamen Verwendung von RASTA-PLP und PI1m Merkmalen wird der niedrigste Medianwert von 6 % erreicht. Allergings wird hier durch die Vorverarbeitung keine Verbesserung mehr erzielt. Insgesamt wird durch die Vorverarbeitung keine signifikante Verbesserung erreicht, da sich die entsprechenden Konfidenzintervalle der Medianwerte überlappen.

5.1.3 Merkmale

Um geeignete Merkmale für die Anwendung der Spracherkennung auf TV- und Filmton zu finden, wird eine sequentielle Vorwärtsselektion der Merkmale [45] durchgeführt. Es werden so lange Merkmale hinzugefügt, bis keine Verbesserung des Fehlers mehr erreicht werden kann. Ausgangspunkt dieser Selektion sollen Merkmale sein, die sich im Bereich der Sprachdetektion bewährt haben: MFCC, PLP und RASTA-PLP.

Abbildung 5.2 zeigt den Fehler der drei Merkmale. Die Abbildung zeigt darüber hinaus die Auswirkung auf den mittleren Fehler, wenn nur die ersten 8 statt der 13 Koeffizienten der vorliegenden Merkmale verwendet werden. Kein Merkmal kann sich deutlich von den anderen abheben. RASTA-PLP mit 13 Koeffizienten weist den niedrigsten Medianwert von 6.6 % auf. Der Unterschied zwischen den Merkmalen fällt gering aus, weil nur Merkmale verwendet wurden, die in [19] besonders gute Erkennungsraten erzielen konnten. Durch die Verwendung nur der ersten 8 Koeffizienten ist keine signifikante Verschlechterung des Medianwertes zu beobachten. Daher wurde aus Gründen des Berechnungsaufwands in den folgenden Experimenten das RASTA-PLP-Merkmal mit nur 8 Koeffizienten benutzt.

[1] In den folgenden Abbildungen wird RASTA-PLP mit der Abkürzung RPLP ersetzt.

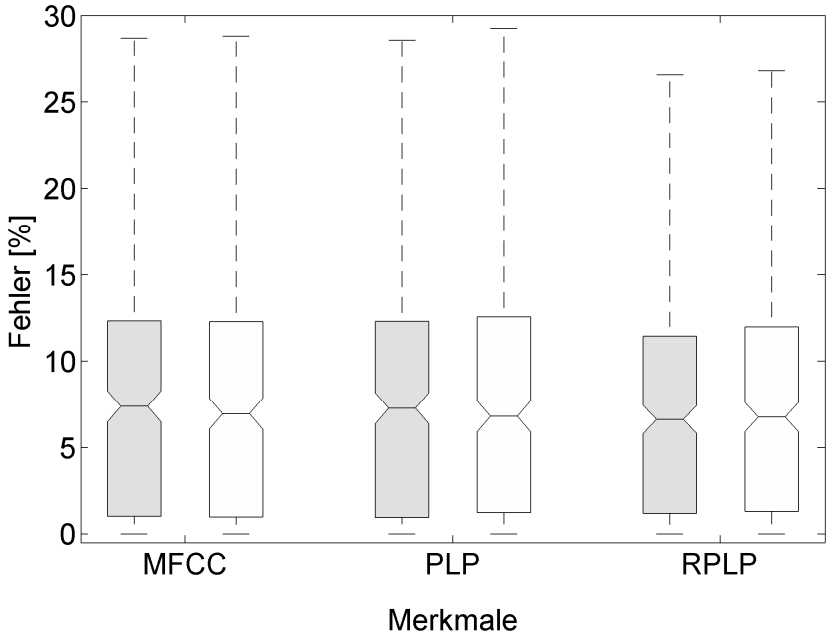


Abbildung 5.2: Vergleich des Fehlers der Items, hervorgerufen durch die Verwendung der ersten 13 (hellgrau) bzw. nur der ersten 8 Koeffizienten (weiß) von MFCC, PLP sowie RASTA-PLP Merkmalen.

Abbildung 5.3 zeigt ein Experiment, welches ermitteln soll, ob durch Hinzufügen eines Stereomerkmals eine Verbesserung des Fehlers erreicht werden kann. Eine signifikante Verbesserung der Medianwerte kann durch das Stereomerkmal nicht erreicht werden, da sich die Konfidenzintervalle überlappen. Der Medianwert erreicht im Fall einer Abtastfrequenz von 24 kHz die beste Verbesserung von 7 % auf 6.6 %. Es kann beobachtet werden, dass das dritte Quartil durch das Hinzufügen des Stereomerkmals geringere Werte annimmt, allerdings auf Kosten von höheren maximalen Fehlern einzelner Items. Da zwischen Medianwerten der Fehler der verschiedenen Abtastfrequenzen kein signifikanter Unterschied vorliegt, wurde die Abtastfrequenz im Hinblick auf den Berechnungsaufwand für alle folgenden Experimente auf 12 kHz herunter gesetzt.

In Abbildung 5.4 sieht man den Fehler, welcher durch das Hinzufügen eines Stereomerkmals zu RASTA-PLP erreicht werden kann. Man sieht, dass tendenziell alle Stereomerkmale eine leichte Verbesserung hinsichtlich des Medianwertes bewirken. Das

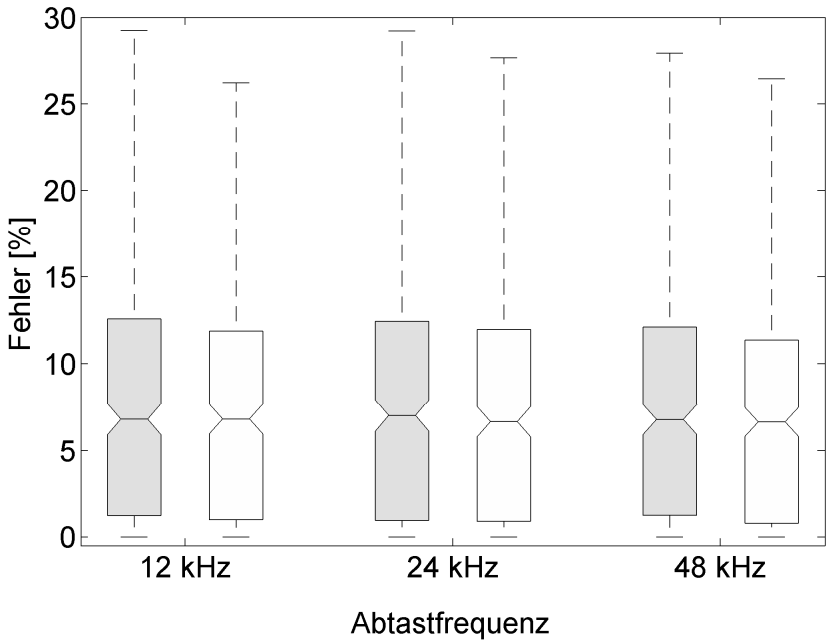


Abbildung 5.3: Fehler der Items, der durch das Hinzufügen des Stereomerkmals ICCm entsteht in Abhängigkeit von der Abtastfrequenz (Grau: nur RASTA-PLP, weiß: RASTA-PLP mit ICCm).

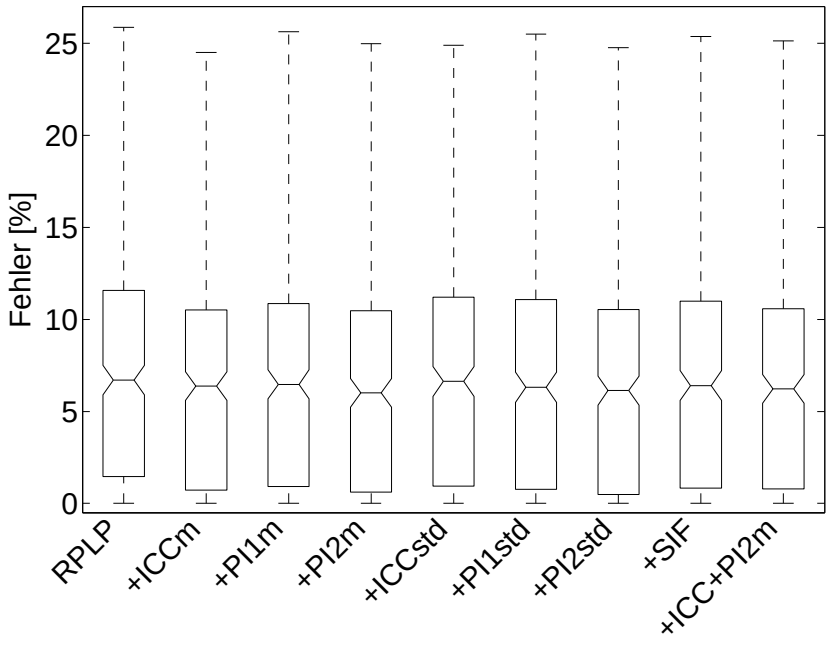


Abbildung 5.4: Fehler, der sich durch das Hinzufügen von Stereomerkmalen zu RASTA-PLP ergibt.

PI2m-Stereomerkmal bewirkt eine Verbesserung des Medianwertes von 6.7 % auf 6 %. Da sich die Konfidenzintervalle der Ergebnisse überlappen, kann jedoch nicht von einer statistisch signifikanten Verbesserung des Medianwertes gesprochen werden.

In Abbildung 5.5 sieht man den Fehler einer sequentiellen Vorwärtsselektion von Merkmalen, ausgehend vom RASTA-PLP Merkmal. Man sieht, dass die Anzahl der Frequenzbänder keinen signifikanten Einfluss auf den Medianwert der Fehler besitzt, da sich die Konfidenzintervalle der Medianwerte überlappen. Durch das Hinzufügen des SF-Merkmals zu RASTA-PLP wird hingegen eine signifikante Verbesserung des Medianwertes von 6.7 % auf 4.4 % für 3 Frequenzbänder bewirkt.

Die Kombination der Merkmale RASTA-PLP und SF dient nun ihrerseits als Ausgangspunkt für die zweite Stufe der Vorwärtsselektion. Es werden die restlichen Merkmale wie in Abbildung 5.5 hinzugefügt. Wie in Abbildung 5.6 zu sehen ist, kann dadurch keine signifikante Verbesserung des Medianwertes mehr erreicht werden.

In Abbildung 5.5 und 5.6 wurde gezeigt, dass durch das Hinzufügen des Spectral Flux

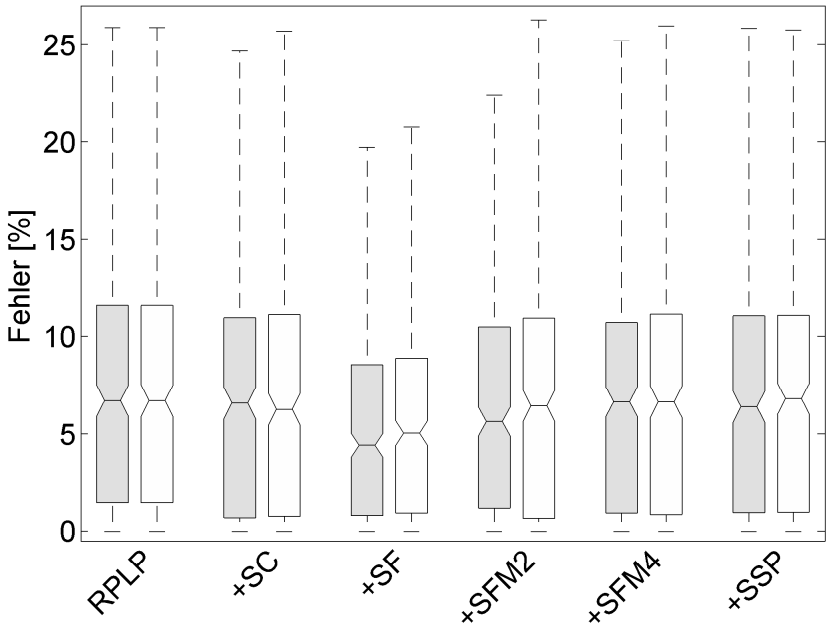


Abbildung 5.5: Vorwärtsselektion ausgehend von RASTA-PLP Merkmal. Hinzugefügt werden spektrale Merkmale (siehe Kapitel 3.1.2), wie SC (Spectral Centroid), SF (Spectral Flux), SFM (Spectral Flatness Measure), SSP (Spectral Spread). Graue Balken: 3 Frequenzbänder. Weiße Balken: 2 Frequenzbänder (oberstes Band wird entfernt).

zu RASTA-PLP eine deutliche Verbesserung des mittleren Fehlers erreicht werden konnte. Nun soll ermittelt werden, ob ein Stereomerkmal zusätzlich zu den beiden Merkmalen ebenso eine Verbesserung des Fehlers bewirkt. In Abbildung 5.7 stellt der erste Balken das Ergebnis von RASTA-PLP dar. Für den zweiten Balken wurde der Spectral Flux zu RASTA-PLP hinzugefügt. Bei allen weiteren Balken handelt es sich um Stereomerkmale, die zur Kombination aus RASTA-PLP und Spectral Flux hinzugefügt wurden. In diesem Fall sieht man, dass die Stereomerkmale nicht in der Lage sind eine weitere Verbesserung des mittleren Fehlers zu erreichen. Alle Stereomerkmale bewirken eine Erhöhung des Medianwertes im Vergleich zu RASTA-PLP mit SF.

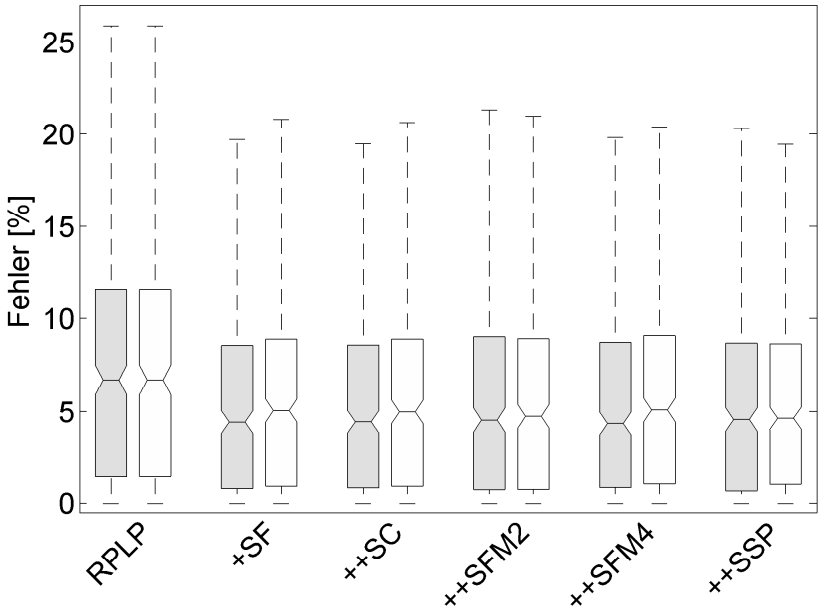


Abbildung 5.6: Zweite Stufe der Vorwärtsselektion, ausgehend von RASTA-PLP und SF.

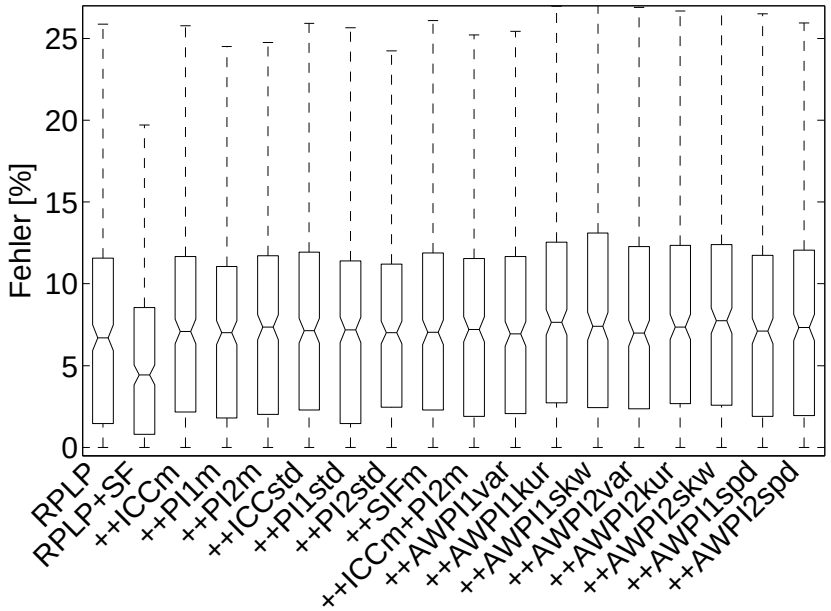


Abbildung 5.7: Fehler der Items durch RASTA-PLP, RASTA-PLP und SF, sowie der jeweiligen Stereomerkmale in Kombination mit RASTA-PLP und SF.

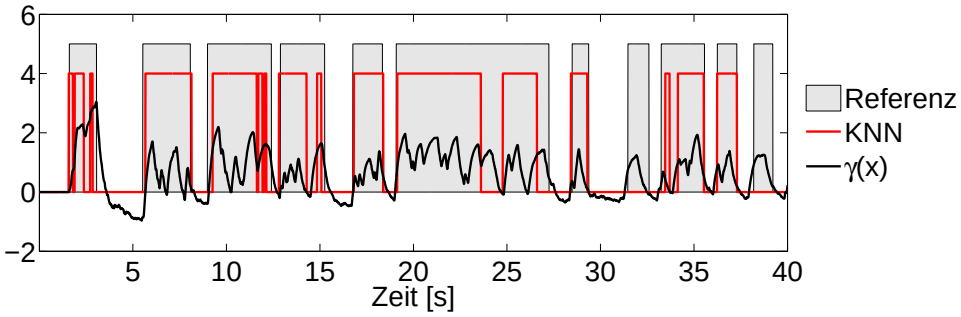


Abbildung 5.8: Verlauf der Differenzfunktion $\gamma(\mathbf{x})$ der modellbasierten Klassifizierung zur Laufzeit. Modellbildung erfolgte mittels Referenz.

5.1.4 Nachverarbeitung

Modellbasierte Klassifizierung zur Laufzeit

Um einen Eindruck zu vermitteln, welche Ergebnisse die Klassifizierung zur Laufzeit liefert, wurde ein Sprecher mit diversen Hintergrundgeräuschen auf die Anwesenheit von Sprache untersucht (siehe Abbildung 5.8). Bei den grau hinterlegten Bereichen handelt es sich um die manuell erstellte Referenzkennzeichnung von Sprache. Die rote Linie steht für die Sprachsegmentschätzung des neuralen Netzwerks und nimmt in der Abbildung aus Gründen der Darstellung für Sprachsegmente den Wert 4 an, während Nicht-Sprache Segmente den Wert 0 besitzen. Die schwarze Linie wird durch die modellbasierte Klassifizierung erzeugt und steht für das Ergebnis $\gamma(\mathbf{x}[m])$ von Gleichung 4.26. Ist $\gamma(\mathbf{x}[m]) > 0$, so ist die Distanz des momentanen Merkmalsvektors $\mathbf{x}[m]$ zum nächstgelegenen Sprachmodell kleiner als zum nächstgelegenen Modell von Nicht-Sprache. Es wird demnach auf die Anwesenheit von Sprache gefolgert. Die Umkehrung gilt für $\gamma(\mathbf{x}[m]) < 0$.

Für Abbildung 5.8 erfolgte die Modellbildung des Klassifizierers zur Laufzeit anhand der Referenzkennzeichnungen. Die Bereiche um 32 s und 39 s werden durch die modellbasierte Klassifizierung zur Laufzeit als Sprache erkannt ($\gamma(\mathbf{x}[m]) > 0$), was dem neuralen Netzwerk alleine nicht gelingt (die rote Linie des neuronalen Netzwerks nimmt den Wert 0 an).

Erfolgt die Modellbildung des Klassifizierers zur Laufzeit nun über die fehlerbehaf-

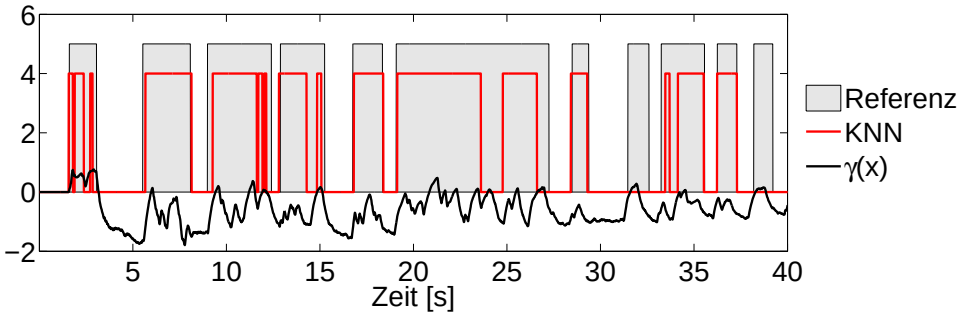


Abbildung 5.9: Verlauf der Differenzfunktion $\gamma(\mathbf{x})$ der modellbasierten Klassifizierung zur Laufzeit. Modellbildung erfolgte mittels Sprachsegmentschätzung.

teten Schätzungen des neuronalen Netzwerks (siehe Abbildung 5.9), nimmt $\gamma(\mathbf{x})$ geringere Werte an wie noch in Abbildung 5.8. Die zuvor erwähnten Bereiche um 32 s und 39 s werden jedoch immer noch großteils detektiert ($\gamma(\mathbf{x}[m]) > 0$). Die modellbasierte Klassifiziererung zur Laufzeit kann also zusätzlich zu den Schätzungen des neuronalen Netzwerks ergänzende Hinweise auf die Anwesenheit von Sprache liefern. Daher wird die modellbasierte Klassifizierung, wie in Abbildung 4.4 gezeigt, mit Hilfe eines Oder-Glieds mit dem neuronalen Netzwerk verknüpft.

Die modellbasierte Klassifizierung wertet den Abstand eines zur Zeit n vorliegenden Merkmalvektors zu Modellen, die Sprache bzw. Nicht-Sprache repräsentieren, aus. Diese Modelle bestehen aus Wahrscheinlichkeitsverteilungen von Merkmalen und werden über Zahlenwerte wie Mittelwert und Varianz beschrieben. Dabei hängt der mittlere Fehler, der durch die Spracherkennung mit modellbasierter Klassifizierung zur Laufzeit erreicht werden kann, auch von der Auswahl der verwendeten Merkmale zur Modellberechnung ab. Abbildung 5.10 zeigt den Fehler, der durch Vorwärtsselektion einer Auswahl von Merkmalen für die Modellbildung, erreicht wird. Die Merkmale Spectral Flatness (SFM2), Spectral Flux (SF), Spectral Spread (SSP), Spectral Centroid (SC) sowie RASTA-PLP (RPLP) werden im ersten Durchlauf getestet. Für SF ergibt sich der niedrigste Medianwert der Fehler mit 7.3 %. Zwischen den Medianwerten der Fehler besteht jedoch kein signifikanter Unterschied. In Folge werden die verbliebenen Merkmale jeweils zu SF hinzugefügt, wodurch sich der Medianwert allerdings verschlechtert.

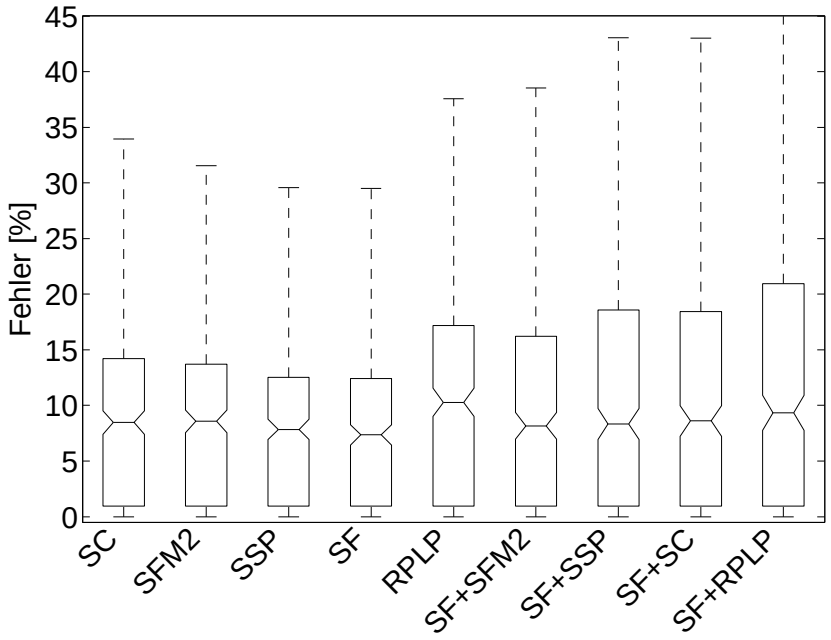


Abbildung 5.10: Modellbasierte Klassifizierung: Vorwärtsselektion von Merkmalen zur Modellbildung.

In Abbildung 5.11 sieht man den Fehler, der sich aus der Verwendung von verschiedenen Distanzmaßen bei der Berechnung von Gleichung 4.26 der Klassifizierung zur Laufzeit ergibt. Der Fehler der ersten zwei Balken ergibt sich aus der Verwendung der Mahalanobisdistanz (siehe Gl. 4.25), mit Hilfe der Varianz (MAHAvar) bzw. der Kovarianzmatrix (MAHAcov). Die restlichen Balken entstanden unter Verwendung der Minkowskidistanz (siehe Gl. 4.24) mit der Ordnungen $P = 1, 2, 3, 4$. Im Diagramm sieht man, dass die Minkowskidistanz erster Ordnung (auch Manhattan-Distanz genannt) den geringsten Medianwert der Itemfehler produziert. Aufgrund der Überlappung der Konfidenzintervalle des Medianwertes kann nicht von signifikanten Unterschieden gesprochen werden.

Die Klassifizierung zur Laufzeit ist abhängig von der Qualität der Schätzungen des ersten Klassifizierers. Dieser gibt die Merkmalsbereiche vor, aus denen die Modelle des zweiten Klassifizierers gebildet werden. Dementsprechend muss es das Ziel sein, Merkmalsvektoren für Modellbildung zu verwenden, bei denen man mit hoher Sicherheit weiß,

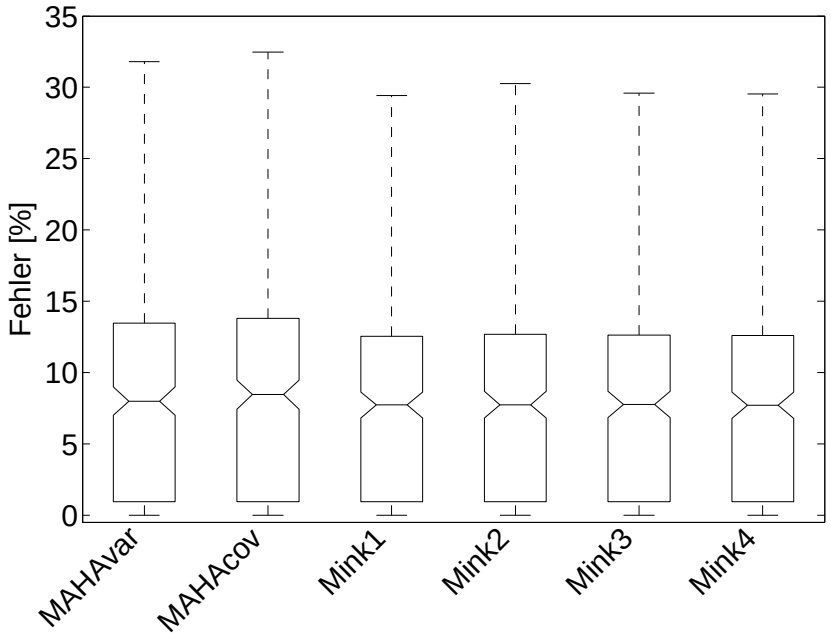


Abbildung 5.11: Modellbasierte Klassifizierung: Untersuchung der Auswirkung verschiedener Distanzmaße auf den Fehler der Items.

dass sie der Klasse Sprache oder Nicht-Sprache angehören. Das Konfidenzmaß des ersten Klassifizierers (hier neuronales Netzwerk) liefert für jede Schätzung einer Klasse einen Wert, der angibt, wie sicher die Schätzung ist. Ein Konfidenzmaß von 0.5 bedeutet, dass es sich sowohl um Sprache als auch um Nicht-Sprache handeln kann. Ein Konfidenzmaß von 1 bedeutet die größte Sicherheit, dass es sich bei einer Schätzung um Sprache handelt. Um die Klassifizierung zur Laufzeit möglichst robust zu machen, wurde ein Schwellwert für das Konfidenzmaß definiert, der bestimmt, welche Merkmalsvektoren für eine Modellbildung berücksichtigt werden. Die Auswertung in Abbildung 5.12 zeigt, dass sich der niedrigste Medianwert von 7.4 % für einen Konfidenzschwellwert von 0.85 ergibt. Der letzte Balken des Diagramms zeigt den mittleren Fehler, der ohne Nachverarbeitung entsteht (7.1 %).

Aus Gründen der Robustheit der Klassifizierung zur Laufzeit wurde eine minimale Modellanzahl gefordert. Sobald jeweils eine bestimmte Anzahl von Sprach- sowie Nicht-Sprachmodellen vorhanden ist, über die Distanzen zum aktuellen Merkmalsvektor be-

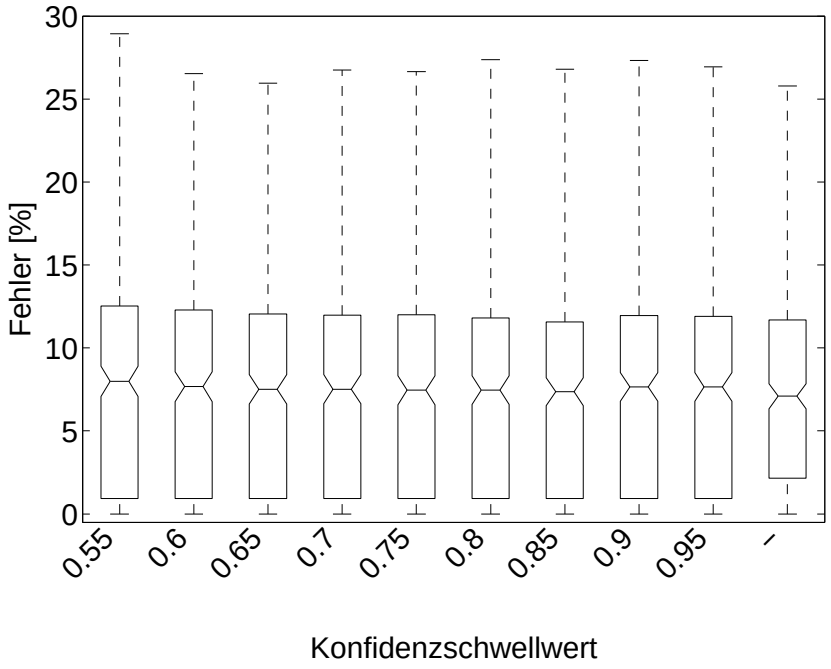


Abbildung 5.12: Modellbasierte Klassifizierung: Veränderung des Schwellwertes des Konfidenzmaßes für die Bildung eines Modells

rechnet werden, wird die modellbasierte Klassifizierung aktiviert. In Abbildung 5.13 wird der beste Fehler für eine minimale Modellanzahl von 7 erreicht (7.2 %). Der letzte Balken steht wieder für den Fehler ohne die Verwendung der modellbasierten Klassifizierung. Zwischen den einzelnen Medianwerten besteht kein signifikanter Unterschied.

Wie man den Abbildungen 5.12 und 5.13 entnehmen kann, wird durch die modellbasierte Klassifizierung zur Laufzeit keine Verbesserung des mittleren Fehlers erreicht. In Abbildung 5.14 stellt die x-Achse den Fehler der durch das neuronale Netzwerk geschätzten Sprachsegmente dar. Hierbei steht jeder Punkt der Abbildung für ein Item (eine Datei der Testmenge, entspricht z.B. einem Filmausschnitt). Die y-Achse zeigt an, wie sich der Fehler der abgebildeten Items durch die modellbasierte Klassifizierung verändert hat. Die Abbildung zeigt, dass sich der Fehler nahezu aller Items, die vor der Verarbeitung einen Fehler bis zu etwa 2 % aufweisen (es handelt sich hierbei um etwa 25 % aller Items), durch die Nachverarbeitung verringert. Die modellbasierte Klassifizierung funktioniert für Items kleinen Fehlers, ist jedoch insgesamt zu abhängig von der

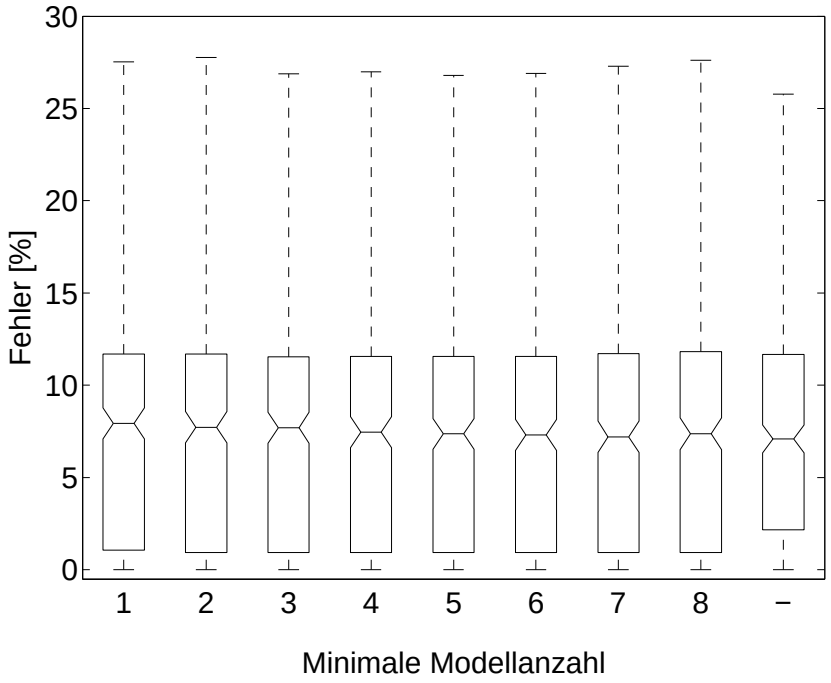


Abbildung 5.13: Modellbasierte Klassifizierung: Auswirkung der Anzahl der geforderten Sprach- bzw. Nichtsprachsegmente auf den Fehler.

Qualität der Schätzungen des neuralen Netzwerks und ist somit nicht robust genug, um eine Verbesserung des mittleren Fehlers über alle Items zu erreichen.

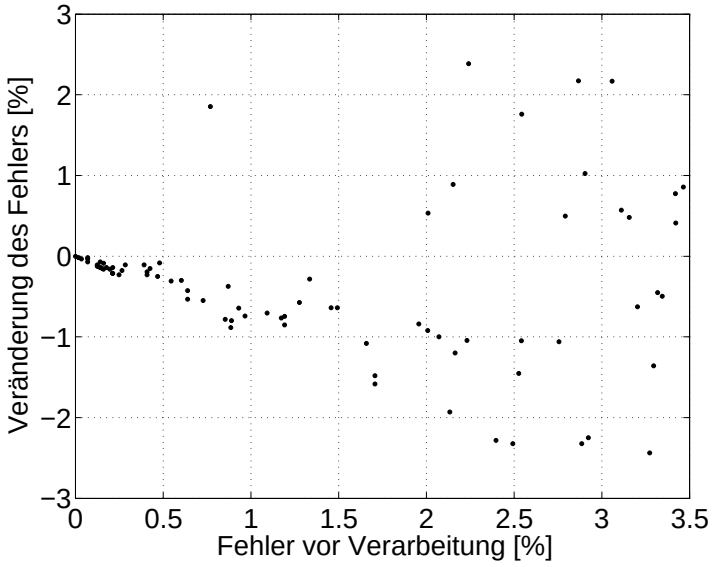


Abbildung 5.14: Veränderung des Fehlers einzelner Items durch die modellbasierte Klassifizierung.

Hüllkurvensegmentierung

Während der Hüllkurvensegmentierung findet eine Verschiebung der durch den Klassifizierer errechneten Sprachsegmentgrenzen statt (siehe Abbildung 4.7). Für den Fall, dass Segmente fälschlicherweise als Sprache klassifiziert wurden, kommt es zu einer Verbreiterung dieser fehlerhaft detektierten Segmente und somit zu einer Erhöhung des Detektionsfehlers. Dieser soll verringert werden, in dem eine Verschiebung der Grenzen nur für Segmente durchgeführt wird, die mit großer Wahrscheinlichkeit der Klasse Sprache angehören. Es wird deshalb gefordert, dass sich Konfidenzwerte der Segmente für die Dauer einer minimalen Konfidenzzeit über einem Konfidenzschwellwert befinden. Segmente, deren Länge sich unter der geforderten minimalen Konfidenzzeit befindet, werden der Klasse Nicht-Sprache zugeordnet. Die Abbildungen 5.15, 5.16 und 5.17 zeigen den Medianwert des Fehlers aller Items, der sich in Abhängigkeit von der minimalen Konfidenzzeit und der Zeit der maximalen Grenzenverschiebung ergibt. Der Konfidenzschwellwert beträgt dabei für die Abbildungen 0.7, 0.6 bzw. 0.5. Die fett gedruckte, durchgezogene Linie steht für den mittleren Fehler der Sprachdetektion ohne die Hüllkurvensegmentierung. Der Medianwert kann durch die Hüllkurvensegmentierung von 7 % auf 5.7 % gesenkt werden. Es wurden die Konfidenzintervalle des Medianwerts aller Ergebnisse untersucht. Dabei konnte keine Signifikante Verbesserung des Medianwertes durch die Hüllkurvensegmentierung festgestellt werden. Die untere Grenze des Konfidenzintervalles ohne Nachverarbeitung lag bei 6.3 %, die kleinste obere Grenze der Konfidenzintervalle der Ergebnisse mit Nachverarbeitung lag bei 6.4 %.

Während sich der mittlere Fehler bei einem Konfidenzschwellwert von 0.7 von 7.3 % bis 5.9 % erstreckt, ist der Fehlerbereich von 6.2 % bis 5.7 % bei einem Konfidenzschwellwert von 0.5 deutlich schmaler. Gerade in Abbildung 5.17 sieht man, dass der Einfluss der maximalen Grenzenverschiebung auf den Fehler im Vergleich zur minimalen Konfidenzzeit relativ gering ausfällt. Die Verbesserung des mittleren Fehlers wird also weniger durch die eigentliche Aufgabe der Hüllkurvensegmentierung, der Verschiebung der Grenzen von Segmenten erreicht. Größer ist hier der Einfluss der Unterdrückung von Segmenten, deren Länge unter der geforderten minimalen Konfidenzzeit sowie unter

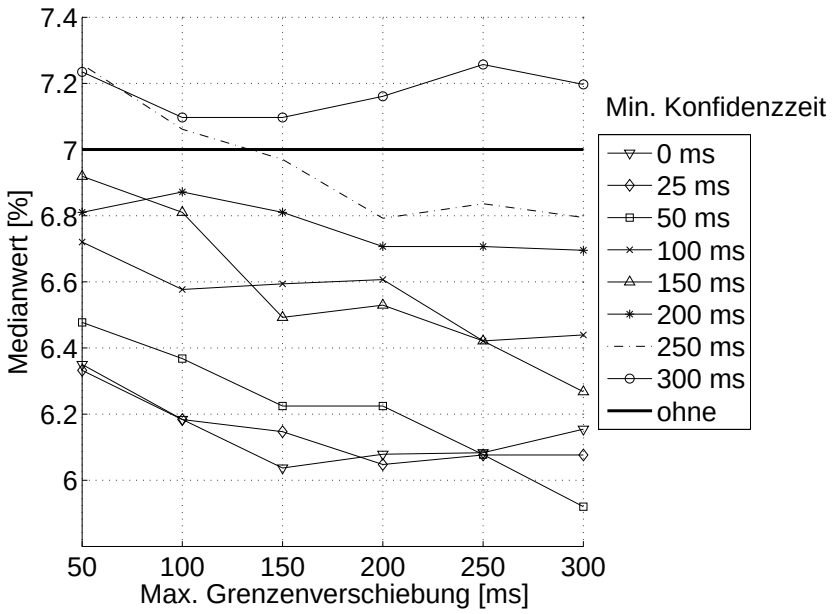


Abbildung 5.15: Hüllkurvensegmentierung: Variation der maximalen Grenzenverschiebung und minimalen Konfidenzzeit. Der Konfidenzschwellwert liegt bei 0.7.

dem Konfidenzschwellwert liegt.

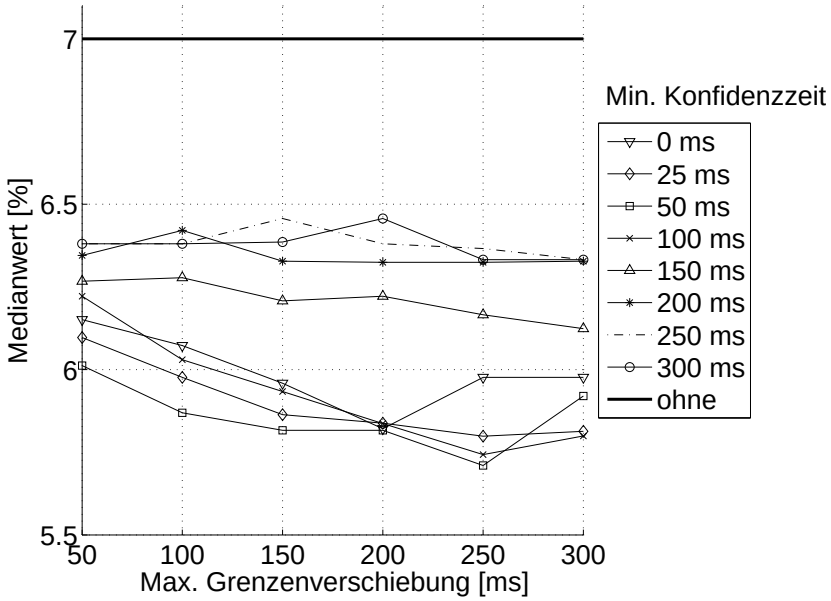


Abbildung 5.16: Hüllkurvensegmentierung: Variation der maximalen Grenzenverschiebung und minimalen Konfidenzzeit. Der Konfidenzschwellwert liegt bei 0.6.

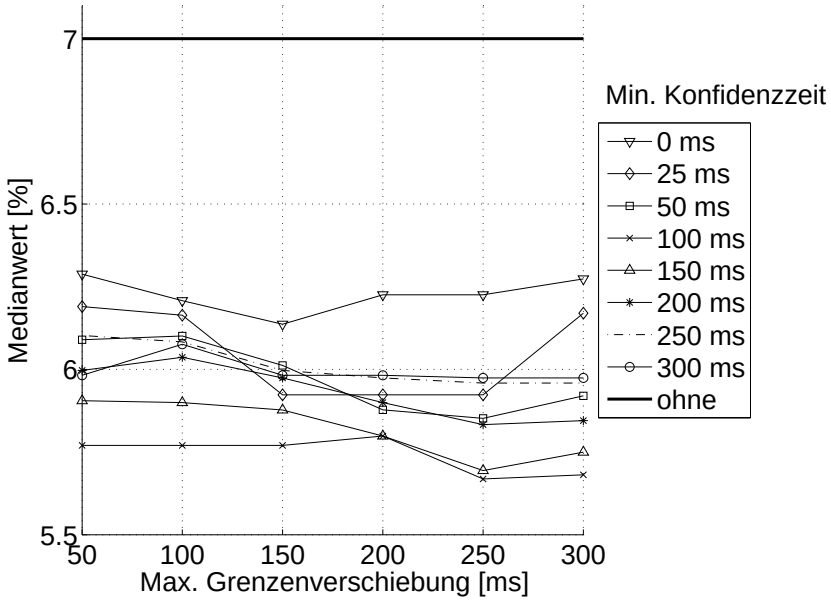


Abbildung 5.17: Hüllkurvensegmentierung: Variation der maximalen Grenzenverschiebung und minimalen Konfidenzzeit. Der Konfidenzschwellwert liegt bei 0.5.

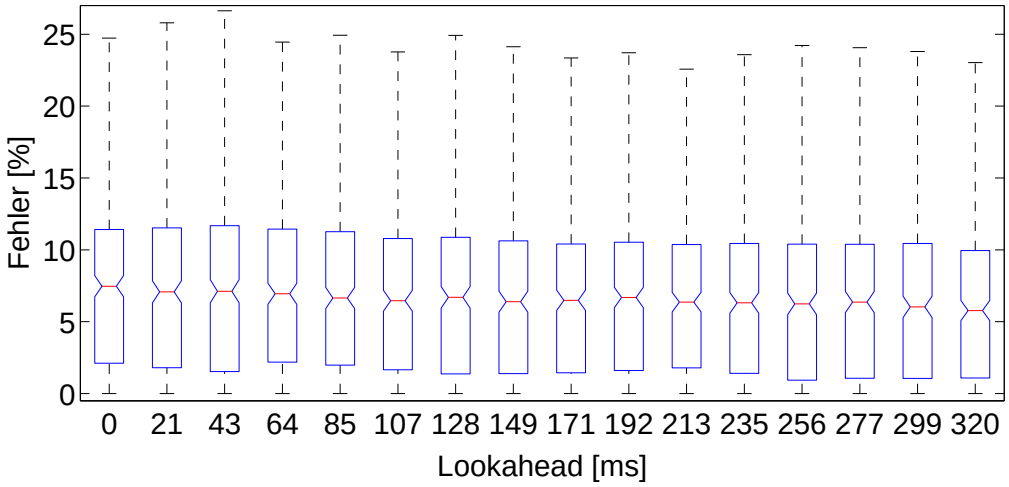


Abbildung 5.18: Einfluss des Lookaheads der Merkmalsextraktion auf den Fehler der Items.

5.1.5 Latenz der Merkmalsextraktion

Abbildung 5.18 zeigt den mittleren Fehler, der sich aus unterschiedlichen Lookaheadwerten ergibt, die durch die Berechnung von Langzeitmerkmalen wie den Sigma Merkmalen (siehe Abschnitt 3.1.3) während der Merkmalsextraktion entstehen. Der Medianwert der Fehler sinkt bei einem Lookahead von 107 ms von etwa 7.4 % auf 6.4 %. Signifikant wird die Verbesserung jedoch erst bei einem Lookahead von 300 ms mit einem Medianwert von 6 %.

5.2 Hörtest

Die im Rahmen dieser Arbeit implementierten Algorithmen wurden durch Hörtests evaluiert und mit dem Stand der Technik des Upmixers verglichen. Die wichtigsten Punkte, die durch den Hörtest evaluiert werden sollen sind:

1. Kann die Forderung nach verbesserter Dialogwiedergabe durch die entwickelten Erweiterungen erreicht werden?

2. Machen sich die Erweiterungen in Form von Artefakten hörbar?

5.2.1 Verwendete Testsignal-Varianten

Dieses Kapitel beschreibt die Erstellung der während des Hörtests verwendeten Klangvarianten eines Testsignals.

Mehrkanal-Studiomischung

Die 5.1 Surround Studiomischung wurde im AC-3 Format [77] von den DVDs der Filme extrahiert. Es lagen 6 diskrete Spuren bei 48 kHz Abtastrate und 16 bit vor. Alle weiteren Klangvarianten gehen in unterschiedlicher Form aus der Studiomischung hervor.

Stereo

Die 6 diskreten Spuren der Studiomischung wurden über einen Downmix auf 2 Stereospuren reduziert. Als Downmix-Methode kam der folgende, durch die ITU für Hörtests empfohlene Referenz-Downmix [78] zum Einsatz. Der Center-Kanal C und die Surround-Kanäle L_S und R_S werden hierbei mit einer Dämpfung von jeweils 3 dB zu den linken L und rechten R Kanälen der 6-Kanal Version addiert:

$$L_0 = 1.00L + 0.71C + 0.71L_S \tag{5.3a}$$

$$R_0 = 1.00R + 0.71C + 0.71R_S \tag{5.3b}$$

Mono

Die Mono-Version eines Testsignals entstand als Downmix ausgehend von der Stereo-Version:

$$M = (L_0 + R_0) * k \tag{5.4}$$

Die Lautstärke der Mono-Version im Vergleich zu den restlichen Klangvarianten wurde über die Konstante k durch den Autor und eine weitere Person nach Gehör angepasst. Abgespielt wurde die Mono-Version ausschließlich über den Center-Lautsprecher.

Music Mode

Die Music Mode Klangvariante liegt im 5.1 Surround Format vor. Sie wurde durch einen Upmix der Stereospur erzeugt. Der Upmixer wurde für alle Klangvarianten, die zum Einsatz kamen im Modus *Spatial* betrieben. Diese Variante repräsentiert den klanglichen Ausgangszustand des Upmixers vor der Diplomarbeit. Die Funktionsweise des Upmixers wird in Abschnitt 3.2.3 beschrieben.

Speech Mode

Die Speech Mode Klangvariante entstand ebenso aus der Stereospur eines Testsignals, die durch den Upmixer in ein 5.1 Surround Format umgewandelt wurde. Der Unterschied zum Music Mode besteht darin, dass der Parameter des Upmixers namens „Center-Integration“ (4.2.2) verändert wurde. Der Speech Mode besitzt ein dominanteren Center-Anteil, was die Wiedergabe von Sprache begünstigt.

Movie Modes

Mit Movie Mode werden die Klangvarianten benannt, die die im Rahmen der Diplomarbeit entstandenen Erweiterungen klanglich repräsentieren. Es wurden 3 Versionen des Movie Modes erstellt, die sich hinsichtlich der entstehenden Latenzzeiten unterscheiden. Wie in Kapitel 4.3.1 beschrieben, können Parameter in den Modulen Sprachdetektion, Nachverarbeitung und Überblendung so eingestellt werden, dass sich die Latenzzeit des

	Movie Mode 1	**Movie Mode 2**	**Movie Mode 3**
Sprachdetektion	0 ms	0 ms	107 ms
Minimale Konfidenzzeit	0 ms	25 ms	50 ms
Pre/Post-Segmentzeit	0 ms	25 ms	50 ms
Überblend-Modus	Echtzeit	Lookahead	Lookahead
Einblendzeit	50 ms	50 ms	150 ms
Ausblendzeit	100 ms	100 ms	100 ms
Gesamtlatenz	≈0 ms	≈100 ms	≈350 ms

Tabelle 5.2: Verwendete Einstellungen der Movie Modes

Gesamtsystems vergrößert. Tabelle 5.2 zeigt die für den Hörtest vorgenommenen Einstellungen der Movie Modes. Für die Berechnung der Latenz des Systems bestehend aus Upmixer und Erweiterungen mit Sprachdetektion muss die Latenz des Upmixers zu den Werten in Tabelle 5.2 addiert werden.

Für die erste Movie Mode Variante wurde eine minimale Latenzzeit gefordert, deshalb findet die Überblendung im Echtzeitmodus (siehe Kapitel 4.2.3) statt und es wird keine Nachverarbeitung eingesetzt. Die zweite Variante verwendet die Hüllkurvensegmentierung (siehe Abbildung 5.17) und die Überblendung findet im Lookahead Modus statt. Bei der dritten Variante wird die Hüllkurvensegmentierung mit größeren Werten verwendet die Sprachdetektion besitzt eine größere Latenz (siehe Abbildung 5.18).

5.2.2 Design der Hörtests

Ein wichtiges Gestaltungskriterium der Mehrkanalerweiterung im Film- und TV-Bereich ist die Wiedergabe von Sprache ausschließlich aus dem Centerkanal [3]. Einerseits war das Ziel der Diplomarbeit, die Sprachwiedergabe des bestehenden proprietären Upmix-Systems hinsichtlich dieses Kriteriums zu verbessern. Andererseits sollen sich die gemachten algorithmischen Erweiterungen nicht störend bemerkbar machen. Diese Forderungen sollen durch einen mehrteiligen Hörtest evaluiert werden.

Hörtest 1: Sprachwiedergabe

Es soll durch diesen Hörtest ermittelt werden, ob die Lokalisation der Sprachwiedergabe des proprietären Upmix-Systems durch die gemachten Erweiterungen verbessert werden konnte. Sprachwiedergabe soll für Filme ausschließlich aus der Mitte, d.h. aus Richtung des Center-Lautsprechers erfolgen. Sitzt man in der optimalen Hörposition für 5.1 Surround Wiedergabe, so besteht zwischen virtuellen und realen Schallquellen aus Center-Richtung bezüglich der Lokalisationsrichtung kein Unterschied. Befindet man sich jedoch in einer nicht idealen Hörposition abseits der Achse, die durch den Center-Lautsprecher und den Mittelpunkt des Kreises verläuft, bewegt sich die Lokalisation einer virtuellen Quelle in Richtung des Lautsprechers, zu dem der geringste Abstand besteht. Maßgeblich verantwortlich ist hierfür der Präzedenz-Effekt (siehe Kapitel 2.2.2). Die Forderung, dass die Sprachwiedergabe aus der Mitte zu erfolgen hat, wird somit für eine nicht optimale Hörposition nicht erfüllt.

Dies bedeutet, dass es besonders für weniger optimale Hörpositionen von großer Bedeutung ist, dass Sprache unter Verwendung des Center-Lautsprechers erzeugt wird. Dieser Hörtest wurde aus diesem Grund nicht in optimaler Hörposition durchgeführt. So kann besser festgestellt werden, welche Klangvarianten den Center-Lautsprecher für die Erzeugung des Sprachsignals mit einbeziehen.

Die optimale Hörposition befindet sich, in Abbildung 2.5 [15] gezeigt, im Mittelpunkt des Kreises. Die gestrichelten Markierungen stellen die ungünstigsten Abhörpositionen dar. Die Abhörposition nahe des rechten Lautsprechers wurde für diesen Hörtest verwendet.

Im Rahmen von Vortests hat sich aufgrund von Rückmeldungen mehrerer Versuchspersonen herausgestellt, dass zwischen den drei Movie Modes im Hinblick auf die zu bewertenden Kriterien lediglich ein geringer Unterschied wahrnehmen ist. Die Zahl der Testsignalvarianten der Hörtests Sprachwiedergabe und Breitenstaffelung wurde daher reduziert. Movie Mode 2 wurde als stellvertretende Klangvariante für alle Movie Modes ausgewählt.

Die Versuchsperson wurde aufgefordert, die Wiedergabe von Sprache im Vergleich

excellent	good	fair	poor	bad

Tabelle 5.3: ITU-Empfehlung für 5-stufige Bewertungsskala

zur Referenz (5.1 Studiomischung) beurteilen. Die Aufmerksamkeit der Versuchsperson wurde dabei besonders auf die Eigenschaften Lokalisationsrichtung und räumliche Ausdehnung von Sprache gelegt.

Als Bewertungsskala wurde eine in 5 Bereiche unterteilte Skala gemäß [78] benutzt (siehe Tabelle 5.3). Die Bewertung kann dabei ganzzahlige Werte von 0 bis 100 annehmen.

Die Testsignale wurden so ausgewählt, dass sie einen hohen Anteil an Sprache aufweisen. Es wurden 6 Testsignale mit 6 verschiedenen Sprecher/innen ausgewählt. Der Anteil an Atmosphäre bzw. Hintergrund-Musik der Testsignale wurde so ausgesucht, dass er einen großen Bereich abdeckt. Die Ausschnitte sind nicht länger als 10 Sekunden. So kann sich die Versuchsperson leicht mit einer Sequenz vertraut machen.

Hörtest 2: Breitenstaffelung

Der zweite Hörtest soll evaluieren, wie gut die Positionen der Schallquellen des Stereo-Referenzsignals im Front-Panorama durch den Upmix-Vorgang erhalten werden konnten. Es soll ermittelt werden, ob das Konzept, zwischen den zwei Klangeinstellungen Movie- und Speechmode zu wechseln, überhaupt notwendig ist. Denn sollte der Speech Mode die Sprachwiedergabe verbessern, aber im Vergleich zum Music Mode keine schlechtere Breitenstaffelung aufweisen, bestünde keine Notwendigkeit zwischen den beiden Klangeinstellungen zu wechseln, da der statische Speechmode ausreichen würde.

Die Versuchsperson soll die Klangvarianten bezüglich der Stereo-Referenz hinsichtlich der Lokalisation von Schallquellen zwischen den Front-Lautsprechern bewerten. Geht es um die Breitenstaffelung des Klangbildes, so ist eine Verschiebung der Quellen in Richtung Center-Lautsprecher unerwünscht und ist abzuwerten. Die Versuchsperson nahm für diesen Test in der optimalen Hörposition Platz. Als Bewertungsskala kommt wieder die Skala abgebildet in Tabelle 5.3 zum Einsatz.

Zur Beurteilung der Darstellung der Breitenstaffelung wurden ausschließlich Musiksignale ohne Sprache herangezogen. Die Schallquellen dieser Signale sollten möglichst das gesamte Panorama zwischen den Front-Lautsprechern ausnutzen. Die Verwendung von Musiksignalen bietet sich für diesen Test an, da die Quellen zudem in den meisten Fällen eine feste Position annehmen. Es wurde darauf geachtet, eine große Bandbreite an Musikstilen für den Test zu finden. Die Ausschnitte wurden so ausgewählt und bearbeitet, dass man sie ohne störende Übergänge in einer Endlos-Schleife anhören konnte. Die Dauer der Ausschnitte befand sich im Bereich von 4 bis 13 Sekunden.

Hörtest 3: Übergänge

Der entwickelte Movie Mode sollte durch Hörtest 1 überprüft werden, ob er eine Verbesserung der Sprachwiedergabe bewirken kann. In Hörtest 3 soll nun erfasst werden, ob die Übergänge, die zwischen den Klangeinstellungen gemacht werden, wahrgenommen werden können.

Hierbei wird die Versuchsperson darüber aufgeklärt, dass der zu evaluierende Algorithmus eine Sprachdetektion verwendet, deren Detektionsergebnisse einen Übergang zwischen zwei Klangeinstellungen des Upmixers veranlassen.

Die Versuchsperson wird darauf hingewiesen speziell auf Übergänge zwischen Sprache und Nicht-Sprache zu achten. Als Klangvarianten werden alle Movie Modes sowie die 5.1 Studiomischung verwendet. Die Studiomischung, bei der keine Übergänge stattfinden, soll hier die Funktion einer verdeckten Referenz einnehmen. Es soll hiermit herausgefunden werden, ob die Versuchsperson die Übergänge tatsächlich hört. Eine originale Referenz, mit der die Klangvarianten verglichen werden sollen, gibt es in dieser Form bei diesem Test nicht. Die Versuchsperson befindet sich während dieses Tests in der optimalen Hörposition.

Die Bewertung soll anhand der in Tabelle 5.4 angezeigten Bewertungsskala vorgenommen werden.

Als Testsignale wurden Filmausschnitte gewählt, die sich im Verlauf der Arbeit im Hinblick auf die Übergänge als kritisch heraus stellten. Die Ausschnitte wurden so gewählt,

Mir fällt nichts auf	Höreindruck C
Mir fällt etwas auf, es stört leicht	Höreindruck B
Mir fällt etwas auf, es stört stark	Höreindruck A

Tabelle 5.4: 3-stufige Bewertungsskala für Hörtest 3: Übergänge

dass möglichst wenig Übergänge stattfanden und der Abstand aufeinander folgender Übergänge möglichst groß war. Es wurden kurze Abschnitte, die eine Dauer zwischen 4 und 7 Sekunden aufwiesen, gewählt. So soll sicher gestellt werden, dass sich die Versuchsperson auf wenige Übergänge konzentrieren kann. Bei der Auswahl der Testsignale wurde auf möglichst unterschiedliche Hintergrundgeräusche und Musik geachtet.

In Vortests stellte sich heraus, dass die Überblendungen von mehreren Versuchspersonen, die in der Entwicklung von Mehrkanalton-Systemen tätig sind, nicht wahrgenommen werden konnten. Der Hörtest wurde daraufhin von zwei ausgebildeten Tonmeistern durchgeführt.

Verwendete Klangvarianten aller Hörtests

In Tabelle 5.5 ist ein Überblick über die in den Hörtests verwendeten Klangvarianten zu sehen. Wenn eine Klangvariante als Referenz eingesetzt wird, muss diese als versteckte Referenz noch einmal durch die Versuchsperson erkannt werden und mit der maximal möglichen Punktzahl bewertet werden.

5.2.3 Durchführung der Tests

Die Wiedergabe der Teststücke erfolgte über hochwertige Studio-Abhörmonitore, angeordnet nach ITU-Standard ITU-R BS.775-1 [15], wie in Kapitel 2.3 beschrieben. Die jeweilige, für den Test vorgesehene Sitzposition wurde durch eine Markierung am Boden gekennzeichnet.

Die Wiedergabelautstärke aller Klangvarianten wurden aneinander angeglichen, sodass kein Testsignal nur aufgrund der Lautstärkeunterschiede besser bewertet werden

	Hörtest 1 Sprachwiedergabe	**Hörtest 2 Breitenstaffelung**	**Hörtest 3 Übergänge**
5.1 Studiomischung	Referenz	✗	✓
Movie Mode 1	✗	✗	✓
Movie Mode 2	✓	✓	✓
Movie Mode 3	✗	✗	✓
Speech Mode	✓	✓	✗
Mono	✓	✓	✗
Stereo	✓	Referenz	✗

Tabelle 5.5: Verwendete Klangvarianten der Hörtests im Überblick

konnte. Die Anpassung der Lautstärke erfolgte durch den Autor und einer weiteren Person nach Gehör.

Die Versuchspersonen sind im Umfeld der Audioentwicklung tätig und absolvieren regelmäßig Hörtests am Fraunhofer IIS. Der Hörtest zur Wahrnehmung von Übergängen wurde zusätzlich mit zwei Tonmeistern durchgeführt.

Die drei Tests sind so konzipiert, dass sie durch eine Versuchsperson innerhalb von 15 bis 20 Minuten durchgeführt werden konnten. Zwischen zwei Tests lag eine Pause von mindestens einem Tag.

Die Versuchsperson konnte zwischen den einzelnen Klangvarianten eines Testsignals ohne wahrnehmbare Zeitverzögerung frei umschalten. Die Klangvarianten wurden dabei, für jedes Testsignal unterschiedlich, zufällig angeordnet. Ein Testsignal lief solange in Endlos-Schleife, bis die Versuchsperson die Bewertung für alle Klangvarianten durchgeführt hat. Darüber hinaus bestand die Möglichkeit sich einen kleineren Ausschnitt des Testsignals mit frei wählbaren Grenzen in einer Schleife vorspielen zu lassen. Bei keinem der drei Hörtests wurde Bildmaterial präsentiert.

Alter	24	25	27	30	31	32	33	36	39
Männlich	1	1	2	1	2	1	2	0	1
Weiblich	0	0	0	0	0	0	0	1	0

Tabelle 5.6: Hörtest 1: Sprachwiedergabe. Anzahl der Versuchspersonen nach Alter und Geschlecht.

Fehler [%]	LethalW.	OBrother	Oceans12_1	Oceans12_2	Payback_1	Payback_2
Mode 1	8.7	3.0	7.3	3.9	6.7	4.1
Mode 2	9.6	3.2	7.1	4.8	6.9	4.3
Mode 3	6.2	5.1	6.3	6.7	5.8	11.2

Tabelle 5.7: Fehler der Sprachdetektion der verwendeten Testsignale in Hörtest 1: Sprachwiedergabe.

5.2.4 Auswertung

Hörtest 1: Sprachwiedergabe

Abbildung 5.19 zeigt die Ergebnisse von 12 Versuchspersonen für alle Testsignale und Klangvarianten. Alter und Geschlecht der Versuchspersonen kann Tabelle 5.6 entnommen werden. Die Bewertungen sind in Form von Mittelwert und 95%-Konfidenzintervall dargestellt. Für die Ergebnisse wurden keine Versuchspersonen aus der Wertung eliminiert.

Die klangliche Ausgangszustand des Upmixers (Music Mode), konnte eine Bewertung im mittleren Bereich erreichen, während die Stereoklangvariante sich im unteren Bereich positioniert. Aufgrund des Präzedenzeffekts war Sprache der Stereoklangvariante aufgrund der Sitzposition nahe des rechten Lautsprechers nicht zentral zu hören und wurde deshalb durch die Versuchspersonen abgewertet.

Die im Rahmen der Arbeit entwickelten Klangvarianten Movie-Mode sowie der Speech-Mode konnten die Sprachwiedergabe hinsichtlich der geforderten Kriterien für alle Testsignale signifikant verbessern. Die beiden Modi erreichten Bewertungen im oberen Be-

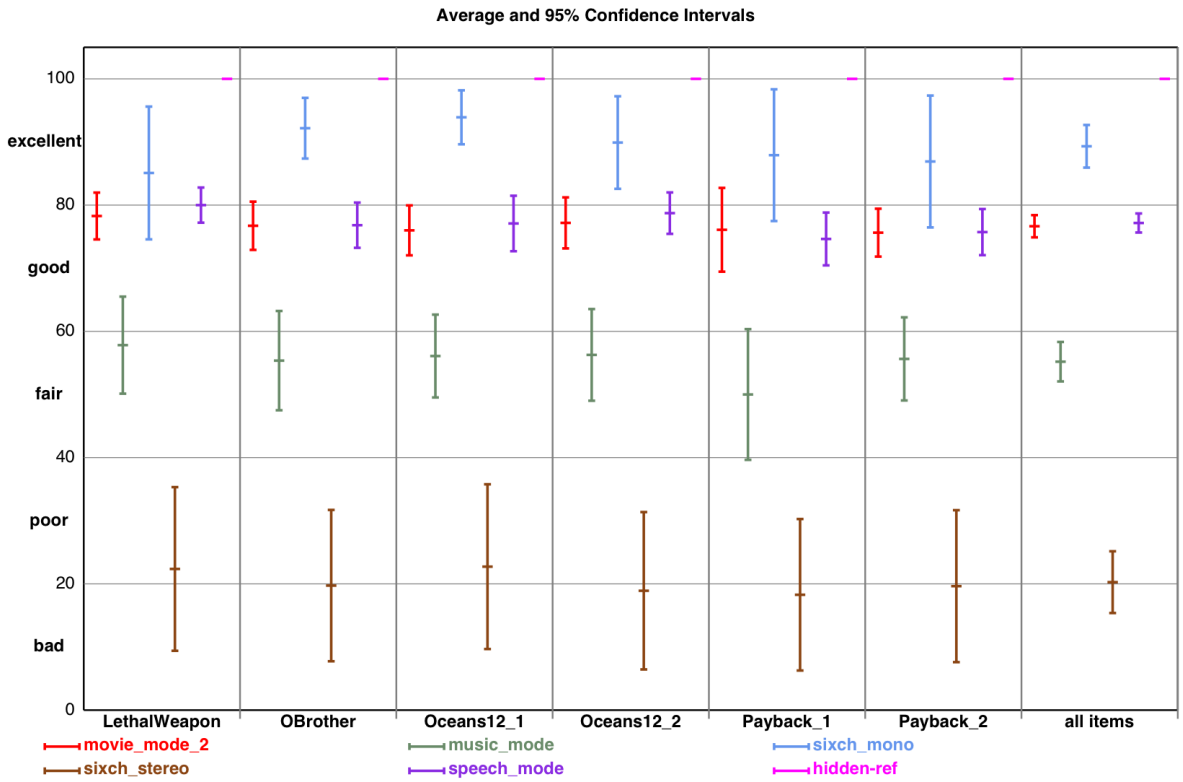


Abbildung 5.19: Ergebnisse von Test 1: Sprachwiedergabe

reich um 75 Punkte und sind somit um mehr als 20 Punkte besser als der Ausgangszustand. Sie werden nur übertroffen von den Klangvarianten Mono und der Mehrkanal-Studiomischung (Referenz).

In Tabelle 5.7 ist der Sprachdetektionsfehler der in Hörtest 1 verwendeten Testsignale abgebildet. Man sieht, dass die Variante Movie Mode 2 des Testsignals *LethalWeapon* mit 9.6 % den schlechtesten Fehler aufweist, der Mittelwert der Hörtestbewertung jedoch der Beste aller Testsignale ist.

Hörtest 2: Breitenstaffelung

Abbildung 5.20 zeigt die Ergebnisse von 11 Versuchspersonen für alle Testsignale und Klangvarianten. Alter und Geschlecht der Versuchspersonen kann Tabelle 5.8 entnommen werden. Die Bewertungen sind in Form von Mittelwert und 95%-Konfidenzintervall dargestellt. Für die Ergebnisse wurden keine Versuchspersonen aus der Wertung elimi-

Alter	24	25	27	28	30	31	32	36	39
Männlich	1	1	2	1	1	2	1	0	1
Weiblich	0	0	0	0	0	0	0	1	0

Tabelle 5.8: Hörtest 2: Breitenstaffelung. Anzahl der Versuchspersonen nach Alter und Geschlecht.

Fehler [%]	Bublee	AirLiquid	Clapton	music_121	music_124	polyhymnia
Mode 1	0	0	0.6	0	0	0.4
Mode 2	0	0	0.4	0	0	0.1
Mode 3	0.9	0	0	0	0	0

Tabelle 5.9: Fehler der Sprachdetektion der verwendeten Testsignale in Hörtest 2: Breitenstaffelung.

niert.

In Tabelle 5.9 sind die Sprachdetektionsfehler der verwendeten Testsignale in Prozent abgebildet. Der maximale Fehler des Movie Mode 2 beträgt nur 0.4 %, d.h. bei reiner Musik findet praktisch keine Fehldetektion von Sprache statt. Dies bedeutet, dass Movie Mode und Music Mode für die vorliegenden Testsignale praktisch identisch sind.

Hinsichtlich der Breitenstaffelung des Klangbildes liegen Music Mode und Movie Mode gleichauf bei etwa 80 Punkten. Der Speech Mode hingegen ist mit einem Ergebnis von etwa 60 Punkten signifikant schlechter bewertet worden als die restlichen Modes.

Betrachtet man die Bewertungen der Upmixklangvarianten von Hörtest 1 und 2, erzielt lediglich der Movie Mode bei beiden Tests gute Bewertungen sowohl für die Sprachwiedergabe als auch für die Breitenstaffelung.

Hörtest 3: Übergänge

Der Hörtest wurde von 5 Versuchspersonen, die in der Audioentwicklung tätig sind, absolviert. Es stellte sich heraus, dass diese Personen keine Übergänge wahrnehmen

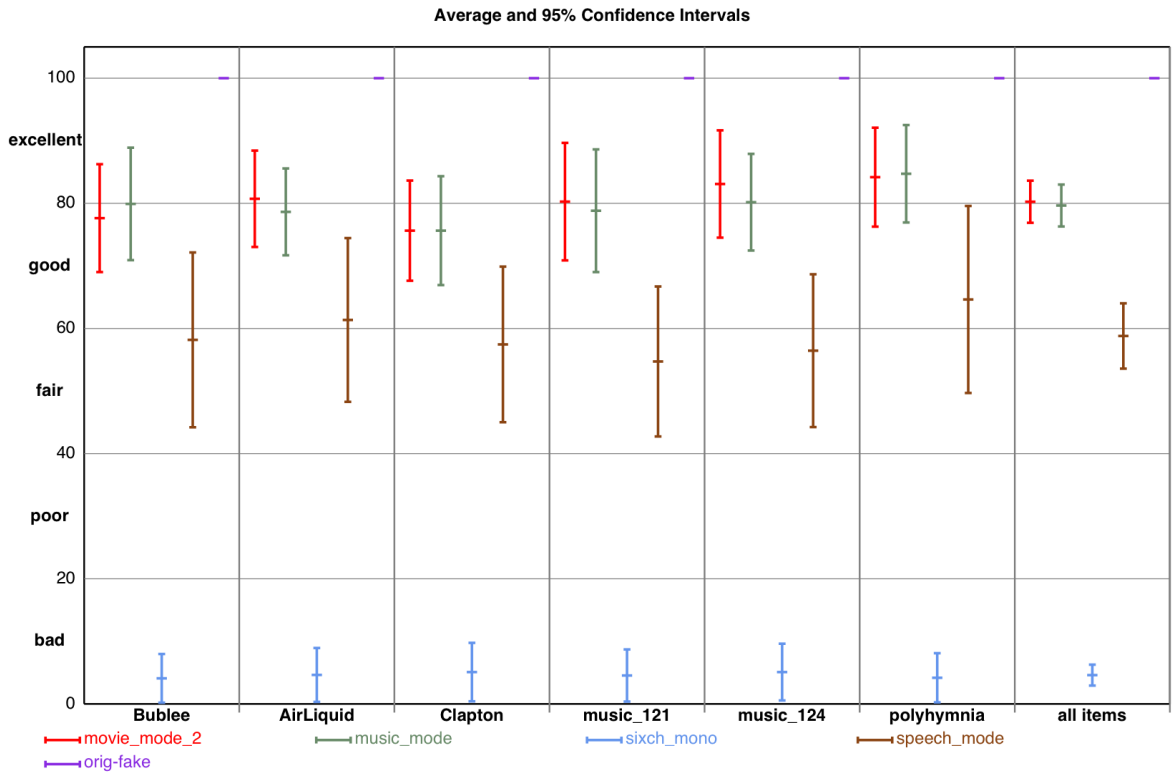


Abbildung 5.20: Ergebnisse von Test 2: Breitenstaffelung des Klangbildes

konnten. Daraufhin wurden zwei ausgebildete Tonmeister gebeten den Hörtest zu absolvieren.

Da eine statistische Auswertung in diesem Fall keinen Sinn macht, werden die Bewertungen in Abbildung 5.21 für jede Person einzeln dargestellt. Das untere Diagramm zeigt einen Audioexperten der die Übergänge unter den Audioexperten am Besten wahrgenommen hat. Dieser gibt an, nur für ein Testsignal Übergänge wahrzunehmen, die ihn aber nur leicht stören. Bei allen verbleibenden Testsignalen nimmt er keine Übergänge wahr.

Tonmeister 1 nimmt für Movie Mode 1 für 3 von 6 Testsignalen keine Übergänge wahr. Bei weiteren 2 Testsignalen nimmt er für diesen Mode Übergänge als nicht störend war. Nur für das Testsignal Payback_1 werden die Übergänge als störend empfunden. Interessanterweise wertet er auch die Surround Studiomischung ab. Als schriftlicher Kommentar gab Tonmeister 1 hierzu folgendes an:

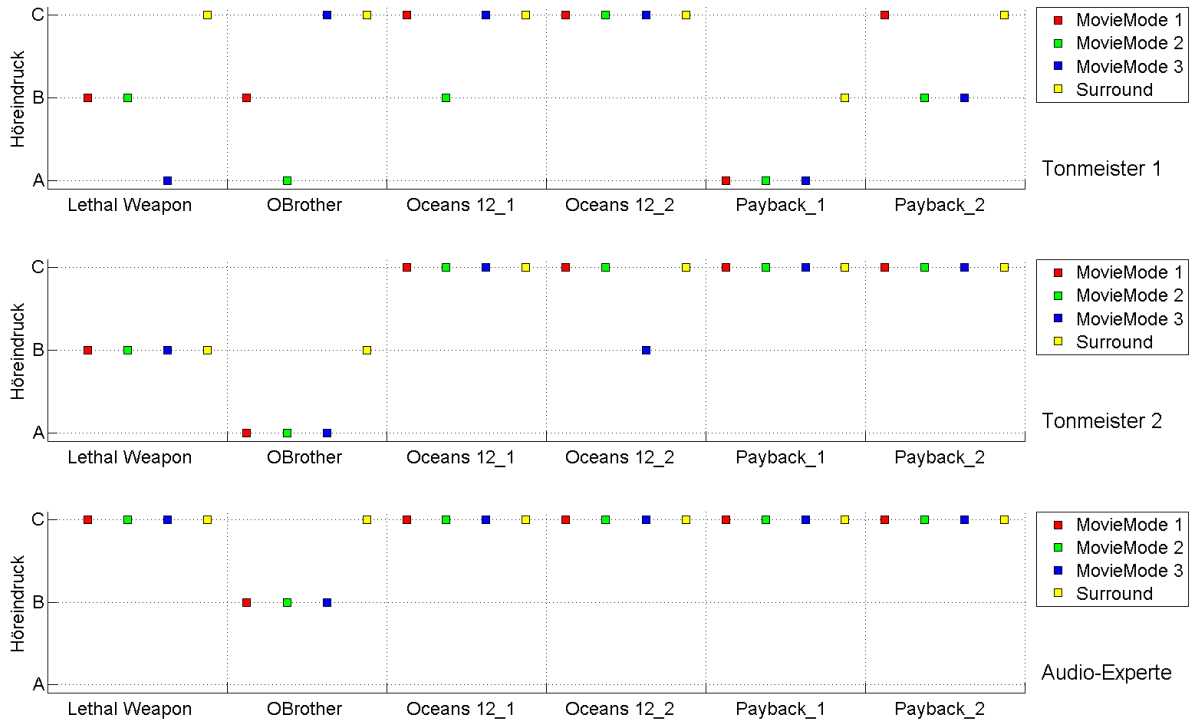


Abbildung 5.21: Ergebnisse von Test 3: Übergänge. Höreindrücke siehe Tabelle 5.4.

„Hier pumpt bei allen Items alles, da kann man keinerlei Aussage machen."

Die von ihm wahrgenommenen dynamischen Artefakte scheinen also schon im Surround-Originalsignal enthalten zu sein, aus dem die Movie Modes in Folge hervorgehen. Die Movie Modes 2 bzw. 3 werden für die ersten beiden Testsignale als störend gewertet. Bei Tonmeister 1 geht Movie Mode 1 als bester Movie Mode hervor.

Betrachtet man mit Movie Mode 1, die Version mit der geringsten Latenz, so nimmt

Fehler [%]	LethalW.	OBrother	Oceans12_1	Oceans12_2	Payback_1	Payback_2
Mode 1	5.2	10.1	7.5	3.6	4.6	6.4
Mode 2	7.0	11.9	7.5	3.8	5.5	7.5
Mode 3	11.8	15.6	10.3	10.1	4.5	7.9

Tabelle 5.10: Fehler der Sprachdetektion der verwendeten Testsignale in Hörtest 3: Übergänge.

Tonmeister 2 in 4 von 6 Testsignalen keine Übergänge wahr. Alle Klangvarianten des ersten Testsignals werden durch Tonmeister 1 als wahrnehmbar, jedoch nicht störend gewertet. Sein Kommentar hierzu:

> „Nach dem zweiten „Hey“ ist ein kleiner Fade zu hören. Da es in allen Stücken ist, bin ich mir nicht sicher ob es im Original selbst auch ist. Sonst fällt nichts auf.“

Die Übergänge aller Movie Modes des zweiten Testsignals werden von Tonmeister 2 als störend wahrgenommen. Er schreibt:

> „Man hört bei allen Items einen kurzen Fade, insbesondere nachdem er „Spielen“ sagt.“

Wieder scheinen die wahrgenommenen dynamische Artefakte schon im Surround-Original vorhanden zu sein. Wahrscheinlich handelt es sich bei den gehörten Übergängen um hörbare Auswirkungen einer Kompression, die sich schon im Originalsignal befindet.

Movie Mode 1 wurde von den Versuchspersonen am Seltensten als störend wahrgenommen.

In Tabelle 5.10 ist der Sprachdetektionsfehler der in Hörtest 3 verwendeten Testsignale abgebildet. Bei beiden Tonmeistern gibt es keinen direkten Zusammenhang zwischen Detektionsfehler und Bewertung der Testsignale.

6 Zusammenfassung

Ein System zur automatischen Mehrkanalton-Erweiterung wurde für die Wiedergabe von TV- und Filmton angepasst. Ein Kriterium von großer Bedeutung war hierbei die klanglich unverfälschte Wiedergabe von Sprache aus dem Centerkanal.

Der im Rahmen der Arbeit entwickelte Ansatz basiert auf der Detektion von Sprache. Auf Grundlage der ermittelten Sprachsegmentgrenzen erfolgt eine Überblendung zwischen zwei Klangeinstellungen des Upmixers. Der klangliche Ausgangszustand des Upmixers diente der Darstellung von allen Passagen des Signals, die keine Sprache enthalten. Eine Klangeinstellung für die Wiedergabe von Sprache wurde mittels einer Anhebung des Centerkanals erstellt.

Zunächst wurde die Sprachdetektion speziell an die Aufgabe der Sprachdetektion in TV- und Filmton angepasst. Die Erweiterungen der Sprachdetektion wurden mit Hilfe einer 10-fachen Kreuzvalidierung über ein Datenset bestehend aus etwa 400 Items zu 60 Sekunden bezüglich ihres mittleren Klassifikationsfehlers bewertet. Die spektrale Gewichtung als Vorverarbeitung des Signals erreichte eine statistisch nicht signifikante Verbesserung des Medianwertes von 6.8 % auf etwa 6.5 %. Das Hinzufügen des Stereomerkmals basierend auf dem Panning Index Centroid zu RASTA-PLP bewirkte eine nicht signifikante Verbesserung des Medianwertes der Fehler von 6.7 % auf etwa 6 %. Durch die entwickelte modellbasierte Klassifizierung zur Laufzeit wurde der mittlere Fehler nicht verbessert. Es konnte jedoch gezeigt werden, dass das Sprachdetektionsresultat von Items, die vor der Nachverarbeitung einen Fehler von weniger als 2 % besitzen, durch die modellbasierte Klassifizierung verbessert wird. Die Hüllkurvenbasierte Segmentierung konnte den Medianwert des Fehlers von 7 % auf etwa 5.7 % verbessern.

Diese Verbesserung ist jedoch nicht statistisch signifikant.

Aus den Sprachsegmenten wird eine Kurve für die Steuerung des Parameters des Upmixers berechnet. Es wurden zwei Varianten zur Berechnung der Steuerkurve entwickelt, die sich maßgeblich hinsichtlich der Latenz ihrer Berechnung unterscheiden.

Die algorithmischen Erweiterungen wurden durch einen dreiteiligen Hörtest evaluiert. Es stellte sich heraus, dass die für die Wiedergabe von Sprache angepasste Klangeinstellung (Speech Mode) im Vergleich zur für Musik angepassten Klangeinstellung (Music Mode) eine signifikant schlechtere Bewertung der Breitenstaffelung erreichte. Umgekehrt erreichte der Music Mode eine signifikant schlechtere Bewertung bei der Darstellung von Sprache. Der im Rahmen der Arbeit entwickelte Movie Mode kombiniert die Vorzüge von Speech und Music Mode und erreicht so bei beiden Bewertungskriterien die beste Bewertung unter den Upmixvarianten.

Übergänge zwischen den Klangeinstellungen von kritischen Testsignalen konnten von mehreren erfahrenen Hörern nicht wahrgenommen werden. Daraufhin wurde der Hörtest von zwei ausgebildeten Tonmeistern absolviert. Diese nahmen die Überblendungen in den meisten Fällen überhaupt nicht, oder als nicht störend wahr.

Die Ergebnisse der Hörtests zeigen, dass es den im Rahmen der Arbeit entwickelten algorithmischen Erweiterungen gelingt, die guten Musikwiedergabe-Eigenschaften des ursprünglichen Systems beizubehalten, gleichzeitig aber die Wiedergabe von Sprache signifikant zu verbessern.

7 Ausblick

Während der durchgeführten Hörtests wurde auf das Einspielen von Bildmaterial verzichtet. Audiovisuelle Phänomene wie der Bauchredner- [8] oder der McGurk-Effekt [79, 80] zeigen, dass die Wahrnehmung von akustischen Signalen durch visuelle Information beeinflusst werden kann. Interessant wäre, wie sich die wahrgenommene Wiedergabequalität des Systems durch das Einspielen von Bildmaterial verändert.

Als das System mit manuell gekennzeichneten Sprachsegmentdaten gesteuert wurde, fiel während informeller Hörtests auf, dass die Übergänge verschiedener Testsignale unterschiedlich stark wahrnehmbar waren. Wenn quantifiziert werden könnte, für welche Signaleigenschaften Übergange deutlich hörbar sind, könnte man darauf beispielsweise durch moderatere Klangeinstellungen des Upmixers reagieren.

Ein direkter Zusammenhang zwischen Fehler der Sprachdetektion und der Bewertung der Testsignale in den Hörtests konnte nicht gezeigt werden. Das im Rahmen der Arbeit verwendete Fehlermaß wird im Bereich der Sprachdetektion oft genutzt, liefert jedoch für die Beschreibung dieser Zusammenhänge nur ungenügende Hinweise. Das Fehlermaß wertet die Anzahl der erkannten Frames im Vergleich zur Gesamtanzahl der Frames aus. Die Wahrnehmung von Sprache durch den Menschen erfolgt jedoch in Segmenten wie den Phonemen. Die Frames besitzen einen unterschiedlich großen Einfluss auf die Übermittelung der Semantik der Sprache. Das Fehlermaß macht zudem keine Aussage darüber, wie gut der Zeitpunkt des Übergangs zwischen Sprache und Nicht-Sprache getroffen wurde. Um die genannten Zusammenhänge genauer überprüfen zu können, bedarf es folglich eines an diese Anwendung angepassten Fehlermaßes.

Literaturverzeichnis

[1] H. Peek, J. Bergmans, J. v. Haaren, F. Toolenaar, S. Stan, and S. G. Stan, "Digital versatile discs," in *Origins and Successors of the Compact Disc*, ser. Philips Research, F. Toolenaar, Ed. Springer Netherlands, 2009, vol. 11, pp. 177–232.

[2] A. Walther, "Mehrkanalerweiterung von Stereo-Audioaufnahmen durch intelligente Upmix-Algorithmen," Diplomarbeit, Technische Universität Ilmenau, 2005.

[3] S. Varga, "Gestaltungskriterien beim automatischen Upmix von Zwei- auf Mehrkanalton im Film- und Fernsehbereich," Diplomarbeit, Fachhochschule Düsseldorf, 2009.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, 2nd ed. Wiley-Interscience, November 2000.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, October 2007.

[6] J. Blauert, J. und Braasch, *Räumliches Hören.* Springer-Verlag Berlin Heidelberg, 2008, ch. 3, pp. 87–121.

[7] L. Rayleigh, "On our perception of sound direction," p. 232, 1907.

[8] R. T. Thurlow WR, "Further study of existence regions for the "ventriloquism effect"," J Am Audiol Soc.1(6):280-6, May-Jun 1976.

[9] J. Blauert, "Sound localization in the median plane," Acoustica 22:205-213, 1969.

[10] K. Wendt, "Das Richtungshören bei der Überlagerung zweier Schallfelder bei Intensitäts- und Laufzeitstereophonie," Ph.D. dissertation, Technische Hochschule Aachen, 1963.

[11] V. Pulkki, "Spatial sound generation and perception by amplitude panning techniques," Ph.D. dissertation, Helsinki University of Technology, 2001.

[12] M. Dickreiter, *Handbuch der Tonstudiotechnik*. Saur; Auflage: 6., 1997.

[13] C. Faller and J. Breebaart, *Spatial Audio Processing: MPEG Surround and Other Applications*. John Wiley & Sons, 2007.

[14] R. Chernyak and N. Dubrovsky, "Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise," in *Proc. 6th Int. Congr. on Acoustics Tokyo*, 1968.

[15] *Multichannel Stereophonic Sound System With And Without Accompanying Picture*, International Telecommunication Union Std. ITU-R BS.775-1, 1994.

[16] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Sound source localization in a five-channel surround sound reproduction system," in *Audio Engineering Society Convention 107*, 9 1999.

[17] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, International Telecommunication Union Std. ITU-R BS.1116-1, 1994-1997.

[18] *A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.*, ITU Std. Recommendation G.729-Annex B, 1996.

[19] C. Uhle, O. Hellmuth, and J. Weigel, "Speech enhancement of movie sound," in *125th AES Convention*, 2008.

[20] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, April 1993.

[21] J. Ramirez, J. Gorriz, and J. Segura, "Voice Activity Detection. Fundamentals and speech recognition system robustness," *M. Grimm, and K. Kroschel, Robust Speech Recognition and Understanding*, vol. 0, pp. 1–22, 2007.

[22] C. Uhle, "Applause sound detection with low latency," in *127th AES Convention*, 2009.

[23] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[24] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.

[25] I. T. Nabney, *Netlab*. Springer, November 2001.

[26] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing (2nd ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[27] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[28] *MPEG-7, information technology - multimedia content description interface - part 4: Audio*, Moving Pictures Expert Group ISO/IEC JTC1/SC29/WG11 Int. Standard 15 938-4, Rev. Final Draft.

[29] P. D. Welch, "The use of fast fourier transforms for the estimation of power spectra," *IEEE Trans Audio & Electroacoustics*, vol. AU-15, pp. 70–73, 1967.

[30] P. Masri, "Computer modelling of sound for transformation and synthesis of musical signals," Ph.D. dissertation, University of Bristol, 1996.

[31] J. Gray, A. and J. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 22, no. 3, pp. 207 – 217, jun. 1974.

[32] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '92.*, vol. 1, 1992, pp. 121 –124.

[33] H. Hermansky, B. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, vol. 10, 1985, pp. 509 – 512.

[34] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357 – 366, aug. 1980.

[35] S. S. Stevens, Je, and E. B. Newman, "A scale for the measurement of the psychological magnitude of pitch," *J. Acoust Soc Amer*, no. 8, pp. 185–190, 1937.

[36] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transfom," *IEEE Trans. Comput.*, vol. 23, no. 1, pp. 90–93, 1974.

[37] C. Uhle, "Persönliche Kommunikation."

[38] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[39] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[40] K. Pearson, "Royal society proceedings, 58, 241," 1895.

[41] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[42] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[43] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, 1998.

[44] S. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, pp. 317–327, 1997.

[45] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[46] T. Holman, *Surround Sound: Up and Running*. San Diego, CA: Elsevier, 2007.

[47] S. Disch, C. Ertel, C. Faller, J. Herre, J. Hilpert, A. Hoelzer, P. Kroon, K. Linzmeier, and C. Spenger, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," in *Audio Engineering Society Convention 117*, 2004.

[48] C. Faller and F. Baumgarte, "Efficient Representation of Spatial Audio Using Perceptual Parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, 2001, pp. 199–202.

[49] C. Faller and F. Baumgarte, "Binaural Cue Coding: A Novel and Efficient Representation of Spatial Audio," in *Proc. ICASSP*, vol. 2, 2002, pp. 1841–1844.

[50] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and Applications," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 520 – 531, nov. 2003.

[51] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: Efficient and compatible coding of multi-channel audio," in *Preprint 116th Conv. Aud. Eng. Soc.*, 2004.

[52] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround-The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *Journal of the Audio Engineering Society (JAES)*, vol. 56, pp. 932–955, 2008.

[53] F. Rumsey, *Spatial Audio*. New York, NY, USA: Focal Press, Oxford, 2004.

[54] S. Weinzierl, *Handbuch der Audiotechnik*. Springer Berlin Heidelberg, 2008.

[55] R. Dressel, "Dolby Surround Pro Logic II decoder principles of operation," 2000.

[56] A. Kraemer, "Circle surround principles of operation," SRS Labs, Inc., 2003.

[57] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc*, vol. 54, no. 11, pp. 1051–1064, 2006.

[58] C. Avendano and J. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002, pp. 121–130.

[59] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," vol. 2, 2002, pp. 1957 –1960.

[60] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2004.

[61] C. Avendano and J. Jot, "A frequency-domain approach to multichannel upmix," *Journal-audio engineering society*, vol. 52, no. 7/8, pp. 740–749, 2004.

[62] V. Pulkki and M. Karjalainen, "Localization of amplitude-panned virtual sources I: Stereophonic panning," *J. Audio Eng. Soc*, vol. 49, no. 9, pp. 739–752, 2001.

[63] V. Pulkki, "Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning," *J. Audio Eng. Soc*, vol. 49, no. 9, pp. 753–767, 2001.

[64] D. Griesinger, "Stereo and surround panning in practice," in *112th AES Convention*, 2002.

[65] B. Bauer, "Phasor analysis of some stereophonic phenomena," *Audio, IRE Transactions on*, vol. 10, no. 1, pp. 18 – 21, jan. 1962.

[66] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc*, vol. 33, no. 5, pp. 314–321, 1985.

[67] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[68] V. Pulkki and M. Karjalainen, "Multichannel audio rendering using amplitude panning [dsp applications]," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 118 –122, may. 2008.

[69] G. Tzanetakis, R. Jones, and K. McNally, "Stereo panning features for classifying recording production style," in *Proc. Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 441–444.

[70] Bronstein, *Taschenbuch der Mathematik*. Harri Deutsch, 2005.

[71] M. Schroeder, T. D. Rossing, F. Dunn, W. M. Hartmann, D. M. Campbell, and N. H. Fletcher, *Springer Handbook of Acoustics*. Springer Publishing Company, Incorporated, 2007.

[72] U. Zölzer, Ed., *DAFX: Digital Audio Effects*. John Wiley & Sons, 2002.

[73] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525–533, 1993.

[74] F. Mosteller and J. Tukey, *Handbook of Social Psychology*. Addison-Wesley, Reading, Mass., 1968, vol. 2, ch. Data analysis, including statistics.

[75] L. Liu and M. T. Özsu, Eds., *Encyclopedia of Database Systems*. Springer US, 2009.

[76] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, pp. pp. 12–16, 1978.

[77] C. Todd, G. Davidson, M. Davis, L. Fielder, B. Link, and S. Vernon, "Ac-3: Flexible perceptual coding for audio transmission and storage," in *96th AES Convention*, 1994.

[78] *Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunication Union Std. ITU-R BS.1534-1, 2003.

[79] H. McGurck and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 246-248, 1976.

[80] J. MacDonald and H. McGurk, "Visual influences on speech perception process," *Perception and Psychophysics*, vol. 24, pp. 253–257, 1978.