Untersuchungen von „Dropout-Concealment“-Algorithmen

Diplomarbeit

durchgeführt von

Hubert Johannes Außerlechner

Institut für Elektronische Musik und Akustik
der Universität für Musik und darstellende Kunst Graz

Begutachter: O.Univ.-Prof. Mag. Dipl.Ing. Dr. Robert Höldrich
Betreuerin: Dipl.Ing. Cornelia Falch

Graz, Mai 2006

Für meine Eltern!

In erster Linie möchte ich meinen Eltern herzlich danken, die mir durch ihre Hilfe und Unterstützung dieses Studium ermöglichten!

Ganz besonders möchte ich mich auch noch bei meinen beiden großartigen Betreuern Frau Cornelia Falch und Herrn Robert Höldrich für die fachliche und organisatorische Unterstützung bedanken. Es war stets zu jeder Zeit ein „offenes Ohr“ für meine Fragen und Probleme vorhanden. Vielen herzlichen Dank!

Dem gesamten Team des IEM danke ich für die angenehme Atmosphäre und ständige Hilfsbereitschaft!

Zuletzt möchte ich mich noch bei meiner Freundin Nina für einen wohltuenden Ausgleich, die tatkräftige Unterstützung und liebevollen Ermutigungen bedanken!

Zusammenfassung

Bei der drahtlosen Übertragung von Audiosignalen können Übertragungsfehler auftreten. Um diesen Ausfällen („dropouts“) entgegenzuwirken, werden verschiedene Verschleierungstechniken („concealment“-Techniken) verwendet. Man unterscheidet zwischen Substitutions- („pattern matching“) und Extrapolationsalgorithmen.

Aufgabengebiet dieser Diplomarbeit ist es, die verschiedenen „concealment“-Methoden auf ihre Anwendbarkeit bzw. Effizienz zu überprüfen. Die Analyse der daraus resultierenden Signale erfolgt mittels eines subjektiven und eines objektiven Audiotests. Beide Tests basieren auf der „ITU-R five-grade impairment scale“ (ITU-R BS. 1284). Mit Hilfe dieser Ergebnisse ist es nun möglich, die Qualität der Algorithmen miteinander zu vergleichen. Aus dem Vergleich lassen sich die Effizienz und im Weiteren Optimierungsmöglichkeiten der Algorithmen feststellen.

Mögliche Erweiterungen bzw. Verbesserungen werden für den Substitutionsalgorithmus realisiert. Die „average magnitude difference function“ und die „zero crossing rate“ sind die Grundlagen dieser Erweiterungsalgorithmen. Anhand der Auswertungen eines erneut durchgeführten subjektiven Audiotests kann der Qualitätsunterschied zwischen dem ursprünglichen Algorithmus und seinen Modifikationen ermittelt und anschließend diskutiert werden.

Abstract

Wireless transmission of audio data is afflicted with transmission errors. To account for these dropouts different concealment techniques have been developed. We discern two different methods: substitution (pattern matching) and extrapolation algorithms.

The aim of this diploma thesis is to investigate different concealment methods, and to estimate their applicability and efficiency. By means of a subjective and an objective test the analysis can be executed. Both tests rest upon the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.1284. With these results it is possible to compare the algorithms amongst each other and to find the most efficient methods. Further on optimizations of the algorithms can be determined.

Two enhancements of the substitution algorithm are implemented. The „average magnitude difference function“ and the „zero crossing rate“ are the basis for these enhancements. After the implementation a further subjective test can be executed. With these results it is possible to compare the quality of the enhanced algorithms and the quality of the original algorithms amongst each other. In addition to that it is possible to find out if the enhancements have really improved the quality of the concealment.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Diplomarbeit lässt sich grob in zwei Kapitel gliedern, den theoretischen und den praktischen Teil: als erstes erfolgt eine Erklärung der Funktionsweise von derzeit aktuellen Verschleierungsalgorithmen. Diese Algorithmen lassen sich in Substitutions- („pattern matching“) und Extrapolationsalgorithmen unterteilen.

Beim „pattern matching“ wird ein kleiner Signalausschnitt, meist direkt vor dem „dropout“ (DO), als Schablone verwendet. Diese Schablone schiebt man über einen festgelegten Beobachtungszeitraum. Durch Vergleichen der Samplewerte der Schablone mit den jeweiligen Werten des Beobachtungszeitraumes (z.B. über eine Korrelationsanalyse) wird darin der Ausschnitt mit der größten Ähnlichkeit zur Schablone ermittelt. Jene Samplewerte, anschließend an diese Folge, werden in den Signalausfall hineinkopiert.

Als Alternative zur Korrelationsanalyse wird einerseits eine Art Grundfrequenzerkennung, basierend auf der „average magnitude difference function“ (AMDF), vorgeschlagen, andererseits werden modifizierte Implementierungen von „pattern matching“-Techniken (z.B. „zero crossing rate“ (ZCR)) untersucht. Wesentliche Kriterien für die Algorithmen sind Schnelligkeit und Recheneffizienz.

Die untersuchten Extrapolationsalgorithmen sind zweistufig aufgebaut; im ersten Schritt erfolgt die Extraktion von Signalinformation aus dem Signal unmittelbar vor dem Ausfall durch die Modellierung eines autoregressiven Prozesses bzw. durch die lineare Prädiktion. Zur Verschleierung des Ausfalls wird im zweiten Schritt das Signal mit Hilfe der geschätzten Parameter resynthetisiert.

Das zweite Kapitel, der praktische Teil, befasst sich mit der Durchführung und Auswertung von subjektiven bzw. eines objektiven Audiotests. Eine Aussage über die Qualität eines Algorithmus’ kann entweder mittels objektiver oder mittels subjektiver Beurteilung getroffen werden. Subjektive Bewertungsversuche, auch Hörversuche genannt, erlauben einer Versuchsperson (VP) die Beurteilung des Grades der Störung eines Audiosignals entlang der digitalen Übertragungsstrecke. Die Hörversuche der vorliegenden Arbeit werden als A-B-A-C Vergleichstests durchgeführt, d.h. der VP werden hintereinander das Originalsignal A, das

Testsignal B, wieder das Originalsignal A und das Testbeispiel C vorgespielt. Eines der beiden Testsignale (B oder C) entspricht dabei jenem Signal, dessen Qualität evaluiert werden soll, das andere ist wiederum das Originalsignal. Die VP bewertet nun beide Testsignale mit Hilfe einer (vorgegebenen) Bewertungsskala in Bezug auf das Originalsignal (vgl. Abb. 1.1).

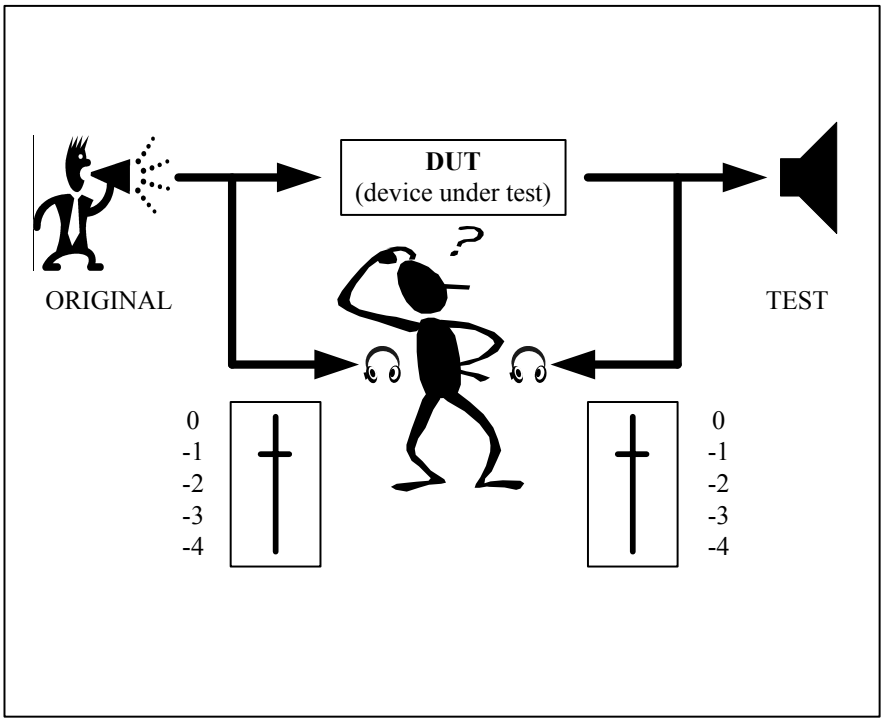

Abb. 1.1: Subjektiver Bewertungstest anhand der 5-teiligen ITU-R Bewertungsskala

Es werden drei Versuchsserien durchgeführt: die ersten zwei dienen der Evaluierung der ursprünglichen Algorithmen, die dritte der Evaluierung der Modifikationen des Interpolationsalgorithmus’. Anhand dieser Ergebnisse ist es möglich, die Algorithmen bzw. ihre Erweiterungen auf Qualität und Effizienz zu überprüfen und zu vergleichen.

Der Ausblick beinhaltet eine kurze Zusammenfassung der durchgeführten Arbeiten bzw. deren Verbesserungen. Weiters wird eine mögliche Erweiterung des Extrapolations-algorithmus’ („long-term, short-term prediction“) behandelt.

2 „Dropout-Concealment“-Algorithmen

Dieses Kapitel beschreibt die Funktionsweise und die Anwendung der zwei am IEM[1] entworfenen Verschleierungsalgorithmen. Nach deren Behandlung wird auf Modifikationen des Interpolationsalgorithmus’ in Bezug auf Recheneffizienz und Qualität eingegangen.
Die erste Erweiterung (siehe beigelegte CD: hja_zcr.m) ist ein „pattern matching“-Algorithmus (vgl. Abschnitt 2.2) und basiert auf der „zero crossing rate“ (vgl. Abschnitt 2.3). Der zweite Algorithmus (siehe beigelegte CD: hja_amdf.m) detektiert unter anderem die Grundfrequenz, kann aber auch als „pattern matching“-Algorithmus angesehen werden, und wird von der Autokorrelationsmethode abgeleitet (vgl. Abschnitt 2.4).

2.1 Allgemeines zu den Algorithmen

Im Rahmen dieser Diplomarbeit werden zwei Verschleierungsalgorithmen mittels subjektivem bzw. objektivem Audiotest evaluiert. Anhand dieser Ergebnisse ist es möglich, Aussagen über die Qualität und Effizienz der Algorithmen zu treffen. Daraus lässt sich wiederum auf Optimierungsmöglichkeiten der Algorithmen schließen. Diese werden in den bestehenden Algorithmus implementiert und erneut mit Hilfe eines subjektiven Audiotests bewertet. Ein Vergleich der daraus resultierenden Ergebnisse liefert erstens eine Aussage über die Qualität der Erweiterungen und der ursprünglichen Algorithmen. Zweitens zeigt sich, ob die Erweiterungen zur Verbesserung oder zur Verschlechterung der Qualität der Verschleierung beitragen.

[1] **I**nstitut für **E**lektronische **M**usik und Akustik
Inffeldgasse 10/3, A-8010 Graz, Austria; Tel.: Tel.: ++43/316/389-3170, Fax: ++43/316/389-3171
http://iem.at/, Email: office@iem.at

2.1.1 Funktionsweise der Verschleierungsalgorithmen

Bei der drahtlosen Übertragung von Audiosignalen können Übertragungsfehler auftreten. Um diesen Ausfällen bestmöglich entgegenzuwirken, werden verschiedene Verschleierungstechniken verwendet. Da die Algorithmen für Echtzeitanwendungen konzipiert werden, ist keine Signalinformation ab dem DO vorhanden (vgl. Abb. 2.1 (a); Signalausschnitt des Violinebeispiels, vgl. Abschnitt 4.2.2). Je nach Verschleierungsalgorithmus gibt es unterschiedliche Ansätze, um den weiteren Signalverlauf vorherzusagen (vgl. Abschnitt 2.1.1.1 und 2.1.1.2).

Allgemein gilt, dass der jeweilige Algorithmus mit Hilfe von Signalinformationen aus der Signalvergangenheit Samples für den Bereich nach dem DO generiert. Dieser Vorgang wird so lange wiederholt, bis wieder das Originalsignal vorhanden ist (siehe Abb. 2.1 (b): die fett gezeichnete Linie kennzeichnet den Bereich der mittels Verschleierungsalgorithmus berechnet wird).

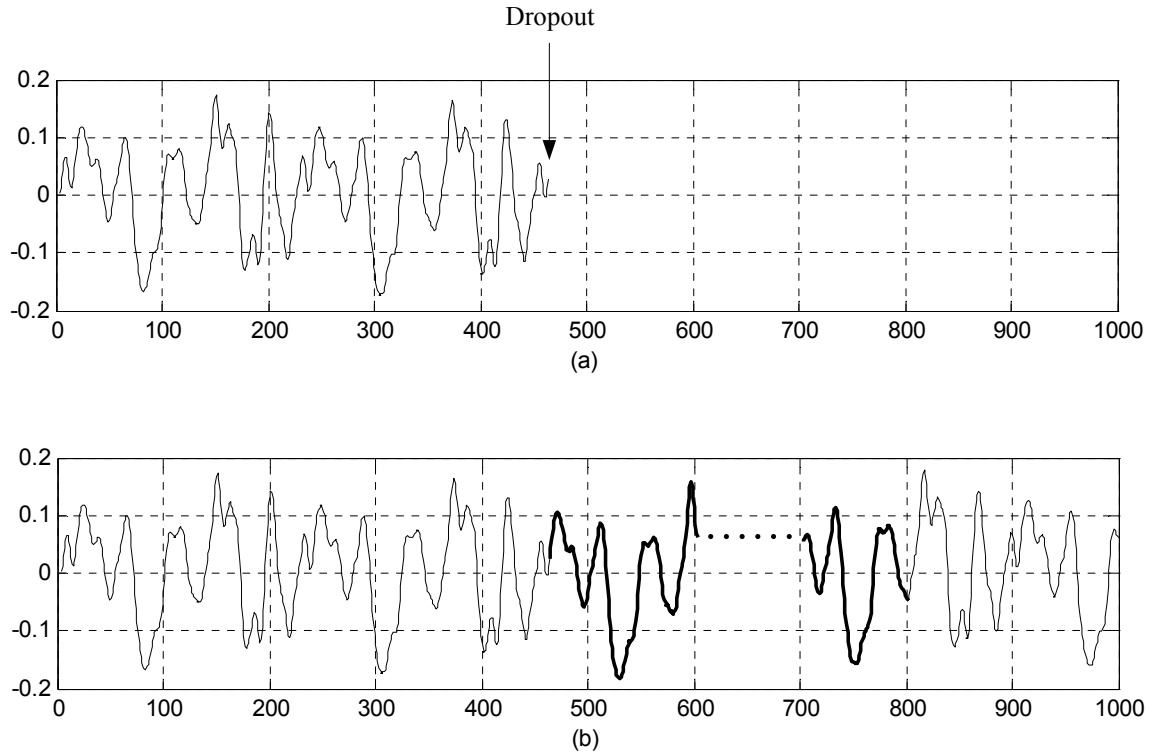

Abb. 2.1: Graphische Darstellung der Funktionsweise von Verschleierungsalgorithmen

2.1.1.1 Lineare Prädiktion

Der erste Algorithmus (Alg1) basiert auf der linearen Prädiktion (LP, vgl. [26]). Man nehme einen Signalausfall an der Stelle M an. Ab diesem Zeitpunkt werden die letzten Samples P[2] zur Schätzung der Pradiktionskoeffizienten verwendet. Mit Hilfe der berechneten Koeffizienten und des vorhandenen Signals lässt sich nun das neue, fehlende Signalsample prädizieren.

2.1.1.2 „Statistische Interpolation“ (Interpolation nach [Vaseghi 1996])

Bei diesem Algorithmus (Alg2) erfolgt die Schätzung der unbekannten Signalsamples mit Hilfe von Signalinformationen vor und nach dem Ausfall. Ein Problem ergibt sich bei Echtzeitanwendungen, da hier die Information nach dem Ausfall nicht vorhanden ist. Abhilfe schafft man sich, indem man mittels Grundfrequenzdetektion vom Ausfallszeitpunkt um eine Periodenlänge in die Signalvergangenheit zurückgeht. Die dadurch gewonnene Signalinformation wird als Information nach dem Signalausfall betrachtet. Mit diesen Daten ist eine statistische Interpolation nun möglich.
Der Vorteil dieses Algorithmus’ gegenüber der linearen Prädiktion liegt darin, dass die Signalenergie während der Signalfortführung konstant bleibt und nicht gegen Null geht.

2.2 Erklärung zum „pattern matching“

Nochmals zur Wiederholung: Beim „pattern matching“ (PM, vgl. Abb. 2.2) wird ein kleiner Signalausschnitt („template“), meist direkt vor dem DO, als Schablone verwendet. Diese Schablone schiebt man über einen festgelegten Beobachtungszeitraum („search window“). Durch Vergleichen der Samplewerte der Schablone mit den jeweiligen Werten des Beobachtungszeitraumes (z.B. über eine Korrelationsanalyse oder die AMDF) wird darin der Ausschnitt mit der größten Ähnlichkeit zur Schablone ermittelt (Abb. 2.2 (b)). Jene Samplewerte, anschließend an diese Folge, werden in den Signalausfall hineinkopiert (Abb. 2.2 (c)).

[2] P … Prädiktionsordnung. Diese bestimmt die Anzahl der Samples, die zur Schätzung der Prädiktionskoeffizienten herangezogen werden.

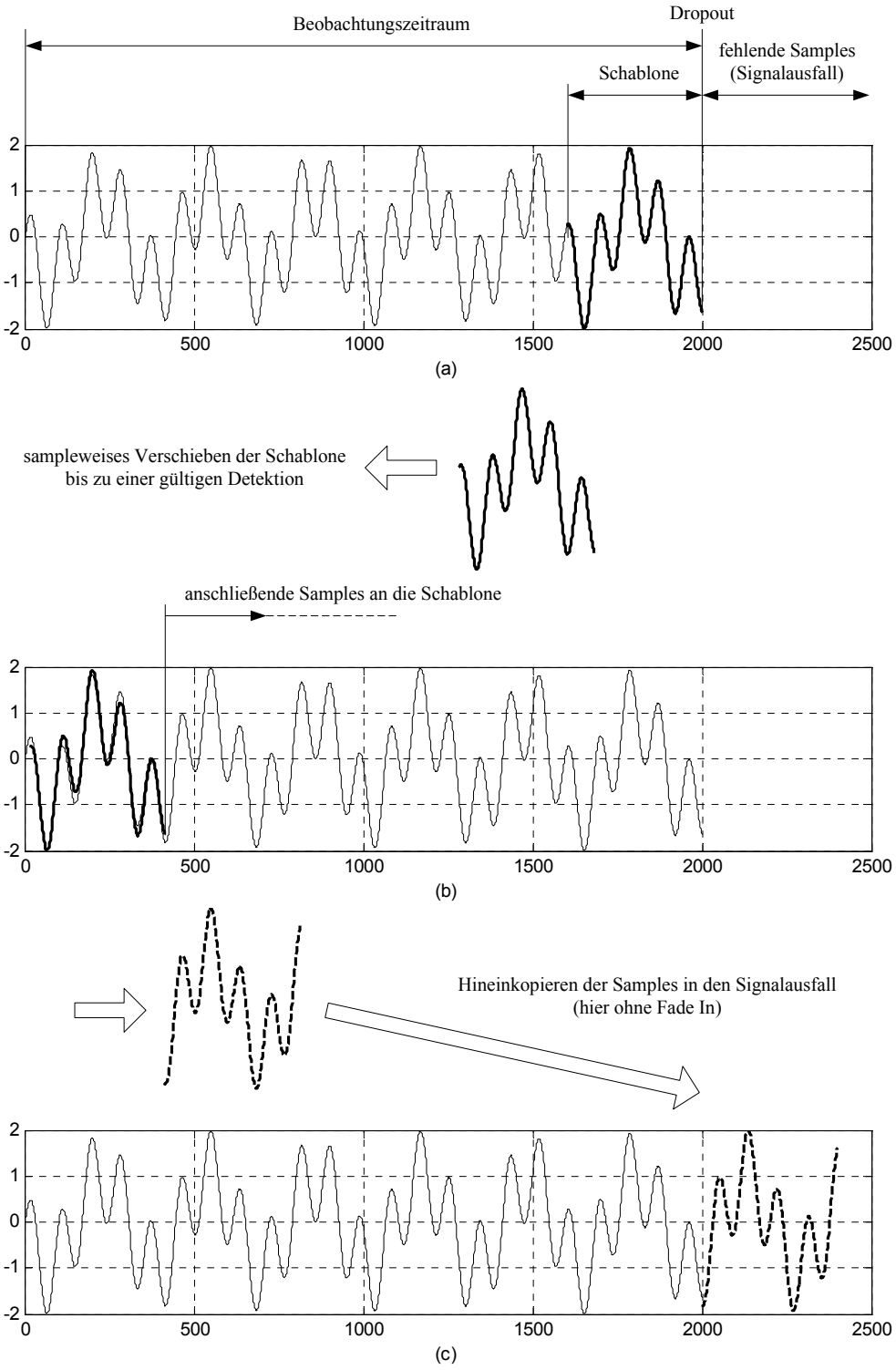

Abb. 2.2: Graphische Darstellung des „pattern matching“-Prinzips

Man kann Grundfrequenz- und im Weiteren PM-Berechnungen entweder im Zeit- oder im Frequenzbereich durchführen. In dieser Arbeit wird auf Grund der Recheneffizienz nur auf Berechnungen im Zeitbereich eingegangen.

2.3 „Zero crossing rate” – ZCR (vgl. [25])

Das Zählen der Nulldurchgänge innerhalb einer festgelegten Zeitspanne ist eine sehr einfache und schnelle Methode um die Grundfrequenz eines Signalausschnitts zu ermitteln. Allerdings funktioniert dies nur bei einigermaßen periodischen Signalen und bei Signalen mit geringer spektraler Dichte. Durch hochfrequente Überlagerungen gibt es mehr Nulldurchgänge, die das Ergebnis verfälschen können.
Die ZCR wird in dieser Arbeit nicht für die Grundfrequenzerkennung verwendet, sondern zu einem PM-Algorithmus umfunktioniert. Betrachtet man den Signalausschnitt in Abb. 2.3, erkennt man, dass man außer der Anzahl der Nulldurchgänge und deren zeitliches Auftreten (Position der Nulldurchgänge) keine brauchbare, recheneffiziente Information für eine Mustererkennung im Zeitbereich zur Verfügung hat.

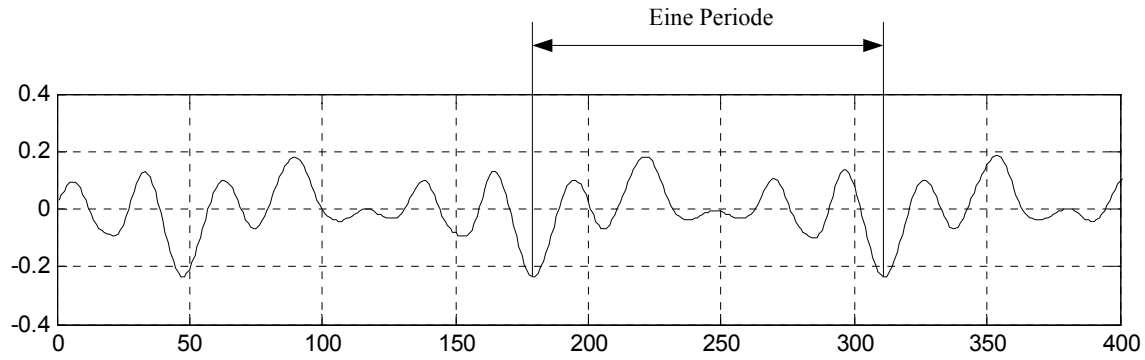

Abb. 2.3: Signalausschnitt (Oboe)

2.3.1 Arbeitsweise des Algorithmus’

Im Folgenden wird der verwendete Algorithmus erklärt. Genaue Programmdetails können dem Programm „hja_zcr.m“ (siehe CD) entnommen werden.

Der Algorithmus besteht im Wesentlichen aus drei Teilen:

- Zählen der Nulldurchgänge; mathematische Kombination: Ort der Nullstellen mit Anzahl der Nullstellen (Division)
- Toleranzbereiche für die folgende Auswahl festlegen
- Aus den detektierten Werten: Ermittlung des Signalausschnitts mit der kleinsten Fehlerenergie

Zuerst werden die Daten (Anzahl der Nullstellen, Ort der Nullstellen dividiert durch die Anzahl der Nullstellen) der Schablone berechnet. Anschließend wird ein Fenster mit der Länge der Schablone sampleweise über den Beobachtungszeitraum geschoben und für jedes Fenster diese beiden Daten ermittelt.

Fensteranzahl = Länge des Beobachtungsraumes - Länge der Schablone + 1

Gleichzeitig erfolgt eine Selektion der Daten nach drei Kriterien (vgl. Abb. 2.4). Der Toleranzbereich bezieht sich dabei immer auf die Daten des Schablonenausschnitts. Die angeführten Kriterien sind ihrer Strenge nach geordnet:

Kriterium 1:

- Gleiche Steigung (positiv/negativ) des ersten Nulldurchganges und gleiche Anzahl der Nulldurchgänge
- Position der Nulldurchgänge innerhalb des Toleranzbereichs ±0.1%

Kriterium 2:

- Anzahl der Nulldurchgänge innerhalb des Toleranzbereichs ±1%
- Position der Nulldurchgänge innerhalb des Toleranzbereichs ±0.1%

Kriterium 3:

- Gleiche Steigung (positiv/negativ) des ersten Nulldurchganges und kleiner gleich dem 2.5-fachen Wert der Anzahl der Nulldurchgänge
- Position der Nulldurchgänge innerhalb des Toleranzbereichs ±0.1%

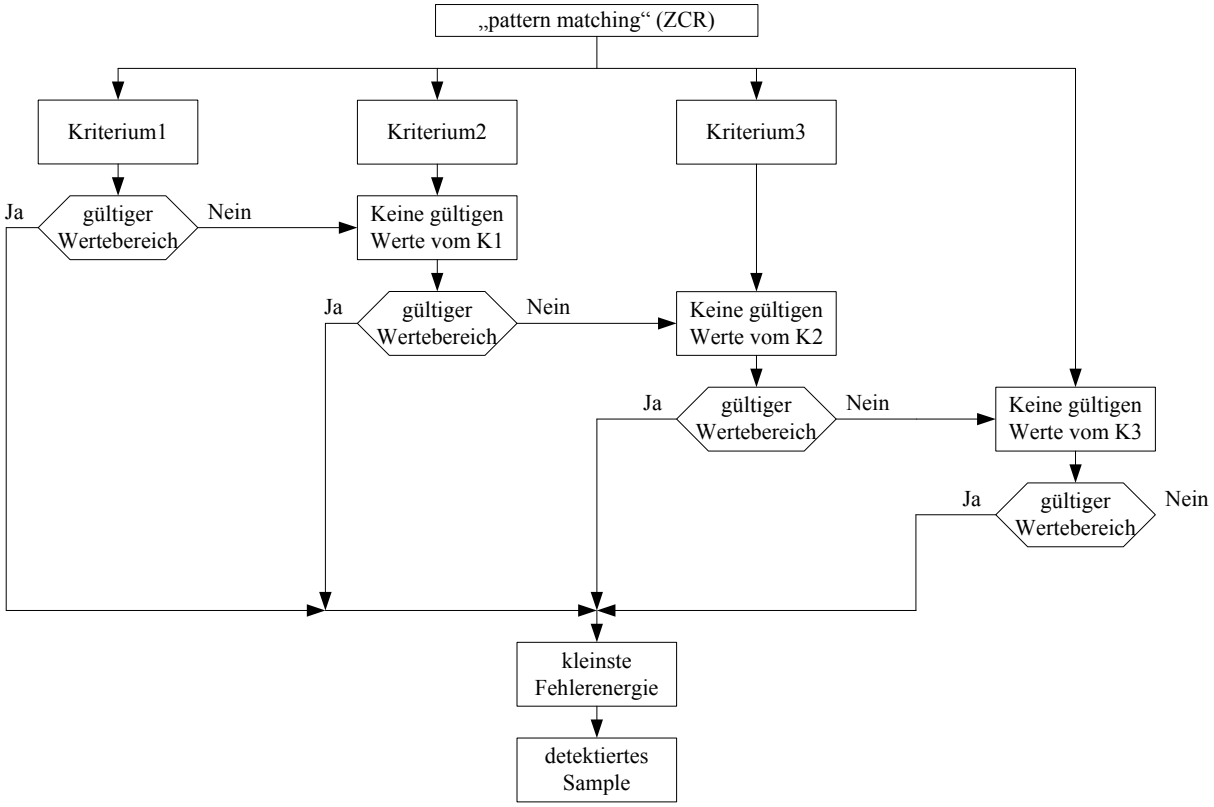

Abb. 2.4: Flussdiagramm, Arbeitsweise des ZCR-Algorithmus’

Untersuchte Zusatzauswahlkriterien:

- Voraussetzung einer Mindestanzahl an Nulldurchgängen innerhalb der Schablone. Bei Unterschreiten dieser Mindestanzahl an Nulldurchgängen wird die Länge der Schablone erhöht. Diese kann aufgrund der Kompatibilität zum Hauptprogramm nicht realisiert werden.
- Man nehme den Fall an, dass die Auswahlkriterien (siehe Kriterien 1-3) kein Ergebnis liefern, d.h. die Detektion erfolgt nur über den kleinsten Fehlerenergieanteil. Der Versuch, eine bessere Detektion über eine Betrachtung der positiven bzw. negativen Energieanteile zu bekommen, liefert die gleichen Ergebnisse.

Für die weitere Auswahl werden die Ergebnisse des ersten Kriteriums bevorzugt. Sollten von dieser Auswahl keine Daten vorhanden sein, wird auf das Kriterium 2 zurückgegriffen, ansonsten auf das Kriterium 3. Sollte auch das letzte Kriterium keine brauchbaren Ergebnisse liefern, werden alle Fenster innerhalb des Beobachtungszeitraums für die Auswahl verwendet. Die Auswahl des Signalausschnitts erfolgt mit Hilfe der kleinsten Fehlerenergie.

$$E_{Fehler} = \left| E_{aktueller\ Bereich} - E_{Schablone} \right| \tag{2.1}$$

Berechnet man die Energie eines um die Zeitachse gespiegelten Signalausschnitts, bezogen

auf den Signalverlauf der Schablone, erhält man als Ergebnis auch eine geringe Fehlerenergie, genauso wie bei zwei fast identischen Signalverläufen (ist nur gültig, wenn $\overline{x}_{Signalausschnitt} \approx 0$). Auch eine getrennte Betrachtung von positivem und negativem Energieanteil liefert in diesem Fall kein besseres Ergebnis.

Ein Fallbeispiel zu einer Fehldetektion ist in Abb. 2.5 (a) dargestellt. Der durch Pfeile markierte Bereich kennzeichnet den detektierten Signalausschnitt. Man erkennt, dass die zwei Funktionen (dünne Linie: Beobachtungszeitraum, dicke Linie: Schablone) nicht übereinstimmen; im Gegensatz zu ihren Energien.

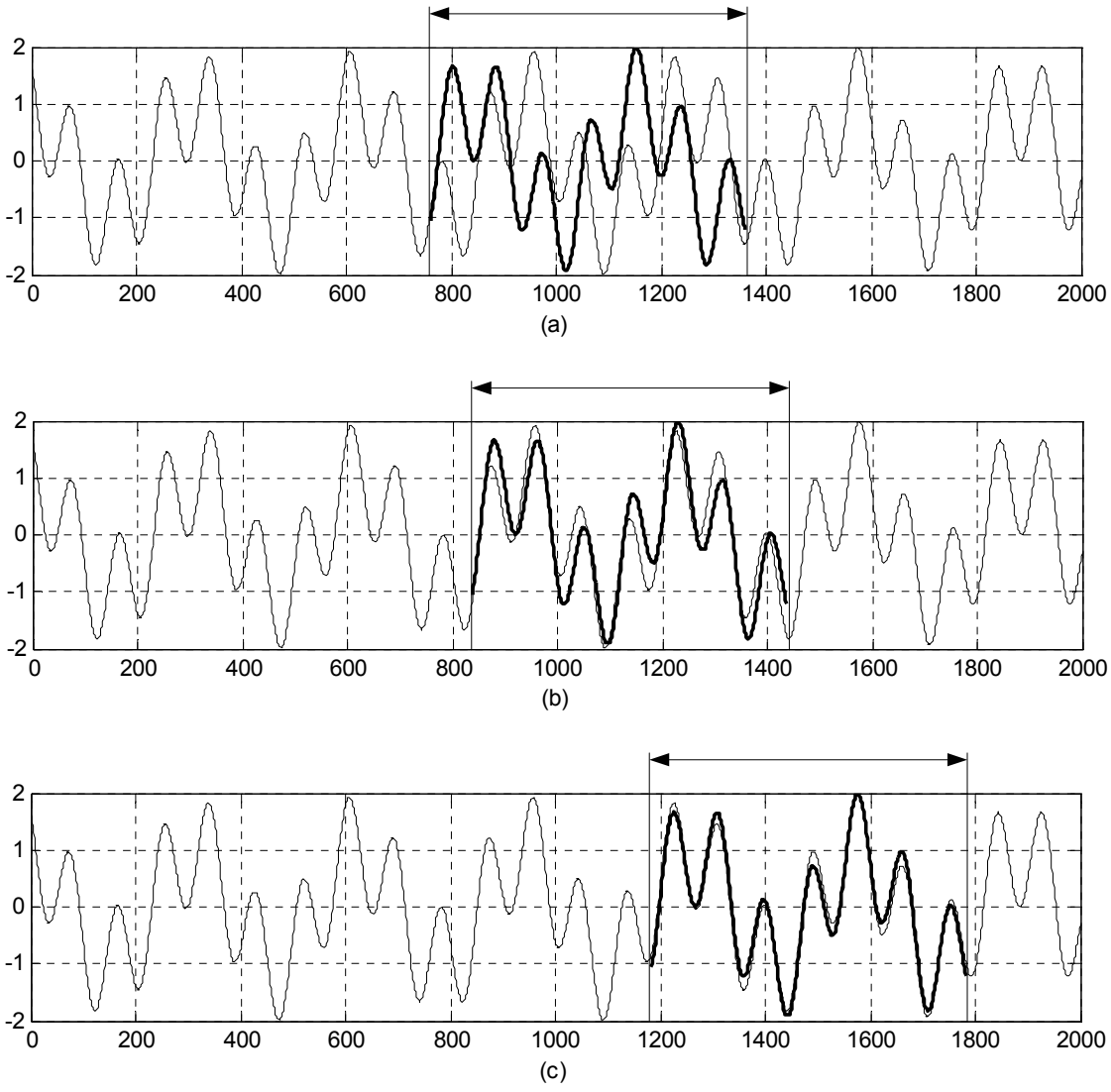

Abb. 2.5: Detektionsbeispiele: (a) fehlerhafte, (b) und (c) gültige Detektion

Mögliche richtige Detektionen sind in Abb. 2.5 (b) und (c) dargestellt. Der Unterschied besteht in der „Anpassung“ (kann auch als Feinabstimmung gesehen werden) der

Nulldurchgänge. Diese Anpassung kann entweder mit dem ersten oder auch mit dem letzten Nulldurchgang durchgeführt werden. Dabei wird die Schablone so weit innerhalb des gefundenen Ausschnitts verschoben, bis jeweils der erste Abb. 2.5 (b) bzw. der letzte Abb. 2.5 (c) Nulldurchgang übereinstimmt. Die Anwendung auf den letzten Nulldurchgang bietet den Vorteil, dass man auf den Fade In zum anschließenden Signal verzichten kann.

Einige Überprüfungen des Algorithmus haben ergeben, dass Berechnungen mit einer Zeitachse mit Offset (Verschiebung der Abszisse um einen positiven oder negativen Wert; wird auch als Schwelle oder „threshold“ (THR) bezeichnet) mehr gültige Detektionen liefern, als mit einer Zeitachse bei der Position $y = 0$. Die Schwelle wird wie folgt berechnet:

$$THR = \frac{\text{Maximalwert des Signalausschnitts}}{2} \tag{2.2}$$

Nach der Berechnung mit dem Offset aus der Formel (2.2) werden die positiven und negativen Energieanteile des detektierten Signalausschnitts ermittelt und mit den Werten der Schablone verglichen. Weisen beide Anteile eine Toleranz kleiner gleich ±0.1% auf, wird der Wert als gültig betrachtet. Bei einer Überschreitung des Toleranzbereichs wird die Berechnung mit einem neuen Offset (vgl. Formel (2.3)) wiederholt.

$$THR = \frac{\text{Maximalwert des Signalausschnitts}}{3} \tag{2.3}$$

Da der ZCR-PM-Algorithmus im Wesentlichen auf dem Zählen und Vergleichen von Werten beruht, liegt sein Vorteil in der Schnelligkeit der Detektion. Geringfügige Abweichungen im Bezug auf die Genauigkeit können deshalb toleriert werden.
Nachteilig erweist sich jedoch sein sehr eingeschränktes Einsatzgebiet. Dieser PM-Algorithmus funktioniert hauptsächlich bzw. am besten für periodische Signale. Eine Fehldetektion kann sowohl bei periodischen als auch bei nicht periodischen Signalen auftreten. Bei weniger periodischen Signalen treten sie jedoch häufiger auf.

2.4 YIN-Algorithmus

Die Bezeichnung des Algorithmus mit „YIN“ (aus der orientalischen Philosophie: „yin“ und „yang“) deutet auf das Wechselspiel von Autokorrelation und ihrer Aufhebung (durch die AMDF bzw. diverse Erweiterungen) hin. Die Entwicklung des Algorithmus besteht aus sechs Schritten (vgl [16]): Autokorrelation – AK, „average magnitude difference function“ – AMDF, „cumulative mean normalized difference function“ – CMNDF, „absolute threshold“, parabolische Interpolation und „Finden des kleinsten Minimums“.

1. Schritt: Autokorrelation – AK

Die AK eines diskreten Signals $x(n)$ kann mit folgenden Formeln berechnet werden:

$$r(\tau) = \frac{1}{p} \sum_{n=0}^{p-1} x(n)x(n+\tau) \tag{2.4}$$

$$r(\tau) = \frac{1}{p} \sum_{n=0}^{p-1-\tau} x(n)x(n+\tau) \tag{2.5}$$

$x(n)$ … Audiosignal (in dieser Anwendung)

n … kennzeichnet den Index einer Folge von Zahlen: $x(n) \in \{x(n); x \in \mathbb{C}, n \in \mathbb{Z}\}$

τ … Zeitabstand (relative Verschiebung)

Für den YIN-Algorithmus wird die Formel (2.4) verwendet. Diese Schreibweise wird auch als „modifizierte Autokorrelation“ oder „Kovarianz“ bezeichnet (vgl. [16]). In Abb. 2.6 sind die beiden Autokorrelationsfunktionen dargestellt (Signalausschnitt aus dem Violinenbeispiel, $p = 900$). Man erkennt deutlich in Abbildung (a), dass die AKF konstant bleibt (vgl. Formel (2.4)), während sie in der Graphik (b) gegen Null strebt (vgl. Formel (2.5)).

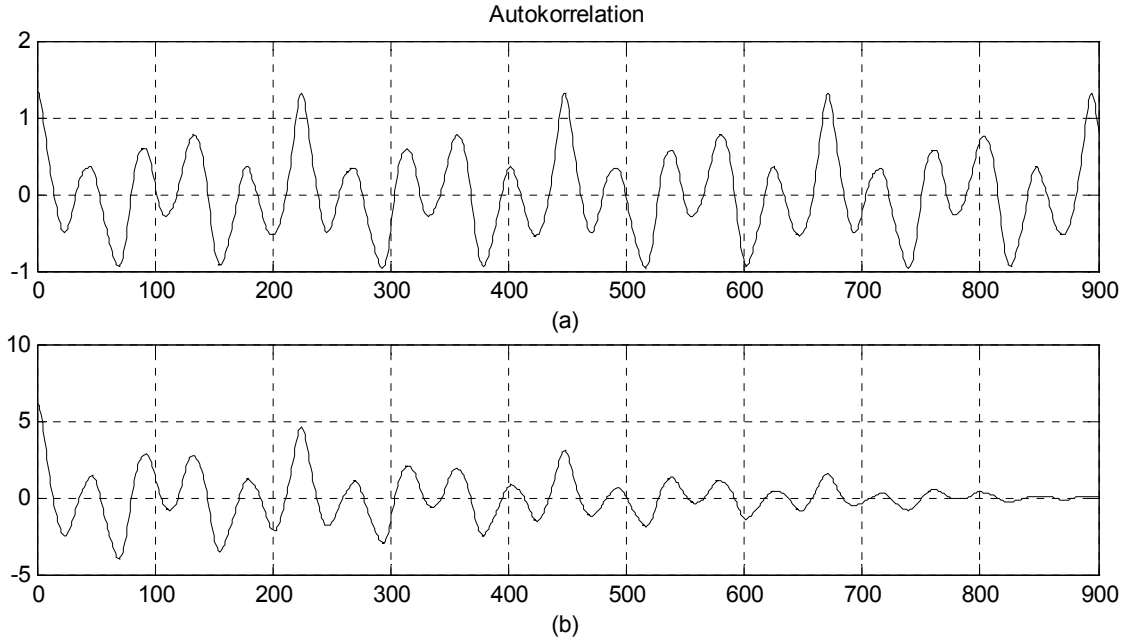

Abb. 2.6: Graphische Darstellung der Autokorrelation: (a) entspricht Formel (2.4), (b) entspricht Formel (2.5)

Über die Datenmatrix X_p des Prädiktionsfehlers (vgl. 6, Formel (6.3)) lassen sich die einzelnen Teile der Autokorrelation erklären (bzw. zusammensetzten):

$$X_p = \begin{pmatrix} x(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x(p+1) & & x(1) \\ \vdots & \ddots & \vdots \\ x(N-p) & & x(p+1) \\ \vdots & \ddots & \vdots \\ x(N) & & x(N-p) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x(N) \end{pmatrix} \tag{2.6}$$

p … kennzeichnet die Ordnung der Prädiktion bzw. der Länge des Prädiktionsfensters bei der Autokorrelation

N … ist der Maximalwert des betrachteten Signalausschnitts: $x(1),\ldots,x(N)$

Diese $(N+p)\times(p+1)$ Rechtecksmatrix lässt sich in drei Teile aufspalten:

$$Xp = \begin{pmatrix} L_p \\ T_p \\ U_p \end{pmatrix} \tag{2.7}$$

wobei die obere Dreiecksmatrix L_p einer $p\times(p+1)$ Toeplitz-Matrix, die Rechtecksmatrix T_p einer $(N-p)\times(p+1)$ Toeplitz-Matrix und die untere Dreiecksmatrix U_p einer $p\times(p+1)$ Toeplitz-Matrix entspricht (vgl. Formeln (2.8)).

$$L_p = \begin{pmatrix} x(1) & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ x(p) & \dots & x(1) & 0 \end{pmatrix}$$

$$T_p = \begin{pmatrix} x(p+1) & \dots & x(1) \\ \vdots & \ddots & \vdots \\ x(N-p) & \dots & x(p+1) \\ \vdots & \ddots & \vdots \\ x(N) & \dots & x(N-p) \end{pmatrix} \tag{2.8}$$

$$U_p = \begin{pmatrix} 0 & x(N) & \dots & x(N-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x(N) \end{pmatrix}$$

Die gesamte Datenmatrix X_p stellt die AK dar, während die Matrix T_p die Daten der Kovarianzfunktion darstellt. In Abb. 2.7 sind die einzelnen Korrelationen dargestellt (Signalausschnitt aus dem Violinenbeispiel, $p = 300$): die gestrichelte Linie kennzeichnet die Korrelation mit der Matrix L_p, die durchgezogene Linie jene mit der Matrix T_p und die punktierte Linie die Korrelation mit der Matrix U_p.

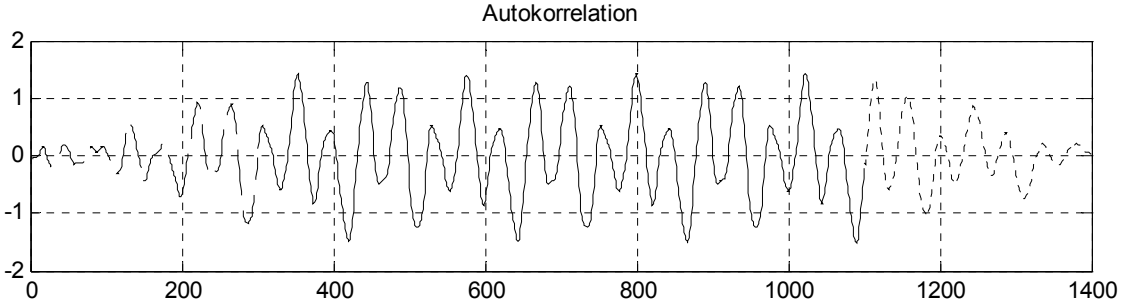

Abb. 2.7: Graphische Darstellung der Datenmatrix Xp

Das charakteristische an der AK ist, dass $r(\tau)$ einen großen Wert annimmt, wenn $x(n)$ ähnlich $x(n+\tau)$ ist. Hat das Signal $x(n)$ eine Periode T, so weist $r(\tau)$ alle $\tau = lT$ $(l \in \mathbb{Z})$ ein Maximum auf. Das globale Maximum der Funktion $r(\tau)$ liegt bei $r(0)$, der nächste Maximalwert bei $r(T)$. Die Grundperiode erhält man somit an der Stelle $\tau = T$.

In manchen Fällen ist es auch möglich, dass ein nachfolgendes Maximum ($\tau \geq 2T$) einen größeren Wert hat wie das Erste. Dies kann einen so genannten „half-pitch“-Fehler zur Folge haben. Der entgegengesetzte Fall, „double-pitch“-Fehler, kann bei einem detektierten Maximum $\tau < T$ auftreten. Eine verbesserte Extremwerterkennung kann mit Schritt 2 durchgeführt werden.

2. Schritt: „average magnitude difference function“ – AMDF

Zur Herleitung der AMDF betrachtet man das Signal $x(n)$ als eine periodische Funktion mit der Periodendauer T (→ invariant für eine zeitliche Verschiebung der Dauer T):

$$x(n) - x(n+T) = 0, \forall n \tag{2.9}$$

Diese Gleichung behält ihre Gültigkeit auch noch nach dem Quadrieren und anschließenden Mitteln über ein Fenster (Länge p):

$$\sum_{n=0}^{p-1} \big(x(n) - x(n+T)\big)^2 = 0 \tag{2.10}$$

Aus der Gleichung (2.10) kann mittels Differenzfunktion eine unbekannte Periode berechnet werden:

$$d(\tau) = \frac{1}{p} \sum_{n=0}^{p-1} \big(x(n) - x(n+\tau)\big)^2 \tag{2.11}$$

Das Ergebnis wird zusätzlich auf $\frac{1}{p}$ normiert. Eine weitere Schreibweise der AMDF stellt Gleichung (2.12) dar:

$$d(\tau) = \frac{1}{p} \sum_{n=0}^{p-1} \big|x(n) - x(n+\tau)\big| \tag{2.12}$$

Je ähnlicher das Signal $x(n)$ und seine Verschiebung $x(n+\tau)$ werden, desto kleiner wird der Wert der AMDF (im Gegensatz zur AK, bei der der Wert ansteigt); d.h. bei einem Signal $x(n)$ mit der Periode T, liefert $d(\tau)$ ein Minimum bei $\tau = T$.

Die graphischen Darstellungen der AMDF zeigt Abb. 2.8. In Diagramm (a) ist die konstante, in Diagramm (b) die fallende Differenzfunktion dargestellt (Zur Erklärung der beiden Versionen kann die gleiche Überlegung wie bei der AK angewandt werden, vgl. Formeln (2.8)).

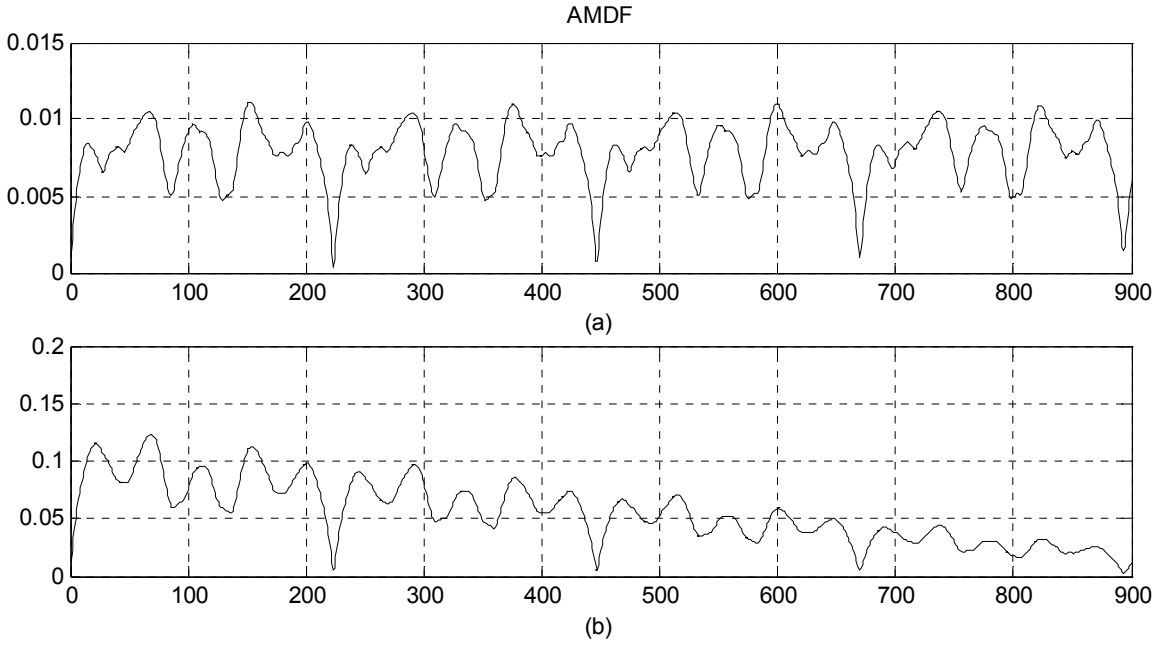

Abb. 2.8: Graphische Darstellung der AMDF

Auch mit Hilfe von Termen der AK kann die AMDF (Formel (2.11)) definiert werden:

$$d(\tau) = r(0) + r_{\tau}(0) - 2r(\tau) \tag{2.13}$$

In dieser Formel entsprechen die ersten beiden Terme Energietermen. Sind diese konstant, ändert sich die AMDF $d(\tau)$ entgegengesetzt der AKF $r(\tau)$, d.h. die Stelle des Minimums der einen bzw. des Maximums der anderen Funktion stimmen überein. Da sich der zweite Term auch mit τ ändert, wird ersichtlich, dass die Maxima von $r(\tau)$ nicht immer mit den Minima von $d(\tau)$ übereinstimmen müssen.

Durch die AMDF wird die Fehlerrate auf 1.95%, von den ursprünglichen 10.0% der AK verbessert (vgl. [16], step 2). Eine Erklärung dafür liefert die Formel (2.4): bei dieser Berechnung ist die AK sehr empfindlich gegenüber Amplitudenschwankungen. Ein zeitlicher Anstieg der Signalamplitude liefert eine ansteigende Amplitude der AK, anstatt einer konstanten (Hess (1983, p.355), vgl. Abb. 2.6 (b)). Dadurch wählt der Algorithmus einen Maximalwert höherer Ordnung aus und verursacht so einen „zu kleinen“ Fehler (eine fallende Amplitude hat genau das Gegenteil zur Folge). Die AMDF hingegen ist immun gegen dieses Problem, da hier Amplitudenänderungen von Periode zu Periode verschiedene

Unähnlichkeiten verursachen. Hess erwähnt auch, dass die Funktion, die Formel (2.5) liefert, unempfindlicher gegenüber Amplitudenschwankungen ist. Die Verwendung der AMDF hat jedoch den zusätzlichen Anreiz, dass diese Funktion ähnlicher dem Signalmodell (vgl. Formel (2.9)) ist und weiters auf die nächsten Fehlerreduktionen überleitet.

3. Schritt: „cumulative mean normalized difference function“ – CMNDF

Die AMDF zum Zeitpunkt Null ist Null ($d(0)=0$) und weist oft infolge ungenauer Periodizität an der Stelle der Periode Werte größer Null auf. Ist der Suchbereich für gültige Werte zu klein, wählt der Algorithmus das Minimum zum Zeitpunkt Null, anstatt dem Minimum beim Zeitpunkt der Periode.

Eine Lösung dieses Problems bietet die CMNDF. Diese berechnet man, indem jeweils der aktuelle Wert der AMDF durch den Mittelwert, bestehend aus den vergangenen Werten und dem aktuellen Wert, dividiert wird (vgl. Formel (2.14) bzw. Abb. 2.9).

$$d'(\tau)=\begin{cases}1, \text{ wenn } \tau=0\\ d(\tau)\Big/\left[\frac{1}{\tau}\sum_{n=1}^{\tau}d(n)\right], \text{ sonst}\end{cases} \tag{2.14}$$

Der Unterschied zwischen CMNDF und AMDF liegt darin, dass $d'(\tau)$ mit dem Wert Eins, anstatt mit Null, beginnt und nur dann Werte kleiner Eins annimmt, wenn $d(\tau)$ kleiner als der Mittelwert ist.

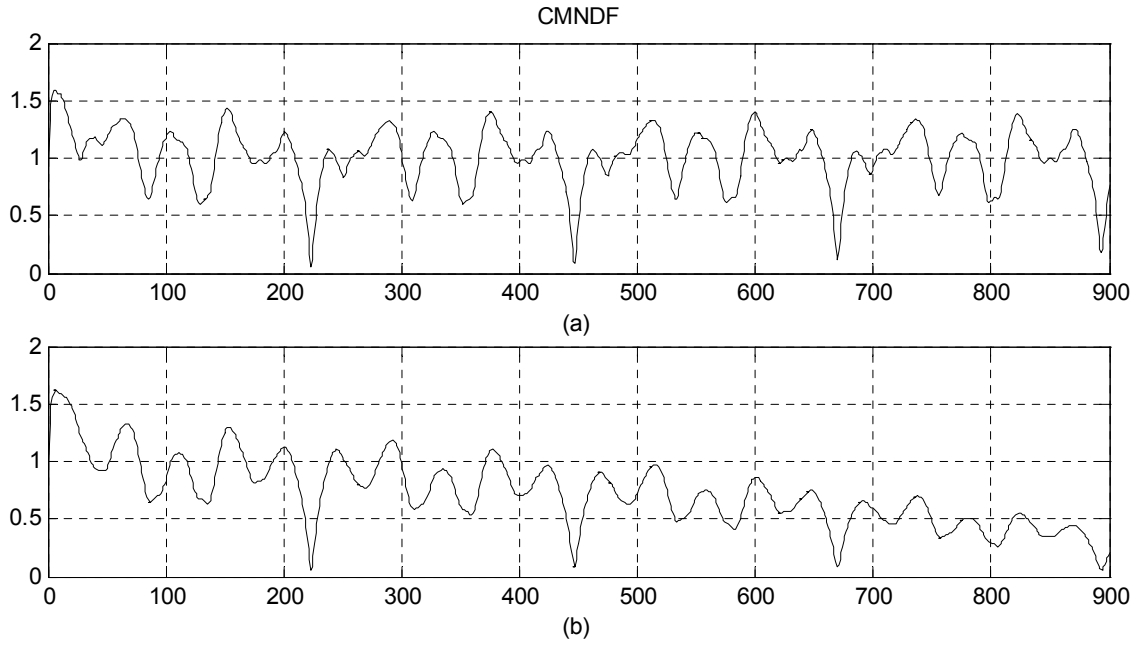

Abb. 2.9: Graphische Darstellung der CMNDF

4. Schritt: Absolute Schwelle („threshold“)

Es kann vorkommen, dass ein Minimum höherer Ordnung kleiner als das Minimum an der Stelle der Periode ist. Fällt dieser Wert in den gültigen Suchbereich für ein Minimum, ist ein subharmonischer Fehler, auch „Oktav Fehler“, das Ergebnis (Die AK ist für solche Detektionen ähnlich anfällig).

Um dieser Fehldetektion entgegenzuwirken, führt man eine absolute Schwelle ein und sucht den kleinsten Wert von τ, der ein Minimum von d' (kleiner als diese Schwelle), liefert. Wird kein passender Wert gefunden, wählt man stattdessen das lokale Minimum.

Die Schwelle stellt ein zusätzliches Auswahlkriterium dar und liefert mögliche gültige Periodenwerte. Sie kann auch als der aperiodische Energieanteil innerhalb des periodischen Signals angesehen werden. Um dies zu verdeutlichen, betrachtet man die folgende Gleichung:

$$2(x(n)^2 + x(n+T)^2) = (x(n) + x(n+T))^2 + (x(n) - x(n+T))^2 \qquad (2.15)$$

Durch Mitteln der Gleichung über ein Fenster W erhält man folgendes Ergebnis:

$$\frac{1}{2W}\sum_{n=0}^{W-1}(x(n)^2 + x(n+T)^2) = \frac{1}{4W}\sum_{n=0}^{W-1}(x(n) + x(n+T))^2 + \frac{1}{4W}\sum_{n=0}^{W-1}(x(n) - x(n+T))^2 \qquad (2.16)$$

Die linke Seite von Gleichung (2.16) entspricht in etwa der Energie des Signals. Beide Terme auf der rechten Seite sind positiv und stellen jeweils einen Teil dieser Energie dar. Der zweite Term ist Null, wenn das Signal eine Periode von T aufweist und wird nicht durch hinzufügen bzw. abziehen von periodischen Teilen zu diesem Zeitpunkt beeinflusst. Dieser Teil kann auch als der aperiodische Energieanteil der Signalenergie betrachtet werden. Bei $\tau = T$ ist der Zähler von Formel (2.14) proportional dem aperiodischen Energieanteil, während der Nenner, Durchschnitt von $d(\tau)$ für $\tau = 0,1...,T$, in etwa dem Doppelten der Signalenergie entspricht. Demnach ist $d'(\tau)$ proportional zu $\frac{aperiodischer\ Energieanteil}{Gesamtenergie}$.

5. Schritt: Parabolische Interpolation

Die vorherigen Schritte arbeiten einwandfrei, wenn die Periode einem Vielfachen der Samplingfrequenz entspricht. Ist dies nicht der Fall, kann das Ergebnis bis zur Hälfte der Samplingfrequenz abweichen. Weiters kann der Fall eintreten, dass große Werte von $d'(\tau)$, die sich nicht in der unmittelbaren Umgebung eines Minimums befinden, den Auswahlprozess für das Minimum stören. Dies führt zu einem größeren Fehler.

Eine Lösung dieses Problems bietet die parabolische Interpolation (polynomiale Interpolation). Durch jedes lokale Minimum von $d'(\tau)$ und seine Nachbarwerte wird eine

Parabel gelegt. Die Ordinate des interpolierten Minimums wird für den Auswahlprozess des kleinsten Minimums, die Abszisse als Periodenwert verwendet.
Die Interpolation von $d'(\tau)$ oder $d(\tau)$ ist rechnerisch einfacher wie ein Überabtasten des Signals.

6. Schritt: Kleinstes Minimum

Die Integrationen in Formel (2.4) und (2.11) garantieren ein stabiles Ergebnis und keine zeitlichen Schwankungen.
Bei nichtstationären Sprachsignalen kann es vorkommen, dass die Schätzung der Periode fehlschlägt. Meistens fällt dies mit großen Werten von $d'(T_t)$ zusammen (T_t kennzeichnet die berechnete Periode zum Zeitpunkt t). Zu einem anderen Zeitpunkt t' stimmt die detektierte Periode, der Wert von $d'(T_{t'})$ ist jedoch kleiner als jener von $d'(T_t)$.
In Schritt 6 wird dieser Fehler berücksichtigt, indem in einem Intervall um den lokalisierten Punkt ein „besseres“ Minimum gesucht wird.

Modifizierter YIN-Algorithmus

Für den PM-Algorithmus werden einige Veränderungen am YIN-Konzept vorgenommen. Da der Algorithmus mittels MATLAB® realisiert wird, können die Schritte 4 bis 6 vereinfacht werden.

Erklärung zu Schritt 4 und Schritt 6:
In Schritt 4 wird zur Verbesserung der Detektion eine Schwelle eingeführt und nur jene Werte betrachtet, die unterhalb dieser liegen. Die MATLAB®-Funktion min(x) macht das Programmieren dieses Schritts hinfällig. Mit Hilfe dieser Funktion wird das lokale (oder auch globale) Minimum eines Signalausschnitts x berechnet. Außerdem kommt es zu keiner merklichen Verbesserung der Ergebnisse durch Einführen einer Schwelle.

Erklärung zu Schritt 5:
Wie schon erwähnt, wird der Algorithmus mittels MATLAB® realisiert. Rechnerisch sind Werte zwischen zwei Samples möglich, praktisch kann das Programm nur mit ganzen Samplewerten rechnen. Wird die parabolische Interpolation angewandt, müssen die Ergebnisse vor ihrer weiteren Verwendung auf ganze Zahlen gerundet werden. Dadurch kann man den Schritt 5 im Vorhinein eingesparen.

AK versus AMDF

Die AMDF ist recheneffizienter als die AK (Subtraktion anstelle von Multiplikation). Weiters ist sie nicht so anfällig für Fehldetektionen. Durch die Erweiterungen des YIN-Algorithmus' sinkt die Fehlerrate zusätzlich.

Eine Version des „Dropout Concealment“ Algorithmus' (Alg2) verwendet zur Detektion das PM-Prinzip. Dafür muss die AMDF in einen PM-Algorithmus ummodelliert werden. Dies erfolgt, indem man zum detektierten Zeitpunkt der AMDF die Anzahl der Samples der Schablone addiert.

Detail über den Programmcode können dem Programm hja_amdf.m (siehe beigelegte CD) entnommen werden.

3 Automatische Audio-Qualitätsmessung

In diesem Kapitel wird auf eine objektive Methode (mittels OPERA™[3]) zur Qualitätsprüfung von Audiosignalen eingegangen, deren Ergebnisse mit jenen der subjektiven Versuche (vgl. Kapitel 2) verglichen werden sollen. Für detaillierte Angaben wird auf [08] und [09] hingewiesen. Die Software OPERA™ (vgl. [10]) wurde von der Firma OPTICOM zur objektiven Evaluierung von Audiosignalen mit sehr geringer Qualitätsverminderung (bezogen auf ein entsprechendes hochqualitatives Vergleichssignal) entwickelt. Als Hauptanwendungsgebiete werden in der Beschreibung die Qualitätsprüfung diverser Komprimierungsalgorithmen, Audiocodecs, Übertragungsstrecken, etc. genannt. Weiters wird angegeben, dass sich die Software zur objektiven Bewertung von Ausfallsverschleierungsalgorithmen eignet. Aufgrund dieser Informationen entstand ursprünglich die Idee, den Schwerpunkt der Untersuchung der Verschleierungsmethoden nicht auf das subjektive sondern das objektive Testverfahren zu setzen. Der subjektive Test soll allgemein eine Einschätzung der Leistung der „concealment“-Algorithmen ermöglichen. Durch den Vergleich der Ergebnisse mit der objektiven Evaluierung wird einerseits die Validierung der gewonnenen Daten erwartet, andererseits kann dadurch in Folge eine ausgedehnte, detaillierte Prüfung mit einer großen Anzahl an Testsignalen vorgenommen werden. Diese wäre mit Hilfe von subjektiven Testreihen wegen des enormen Arbeits- und Zeitaufwands nicht möglich. Wie in Abschnitt 3.4 näher erläutert, erfüllt die Verwendung der Software jedoch nicht die entsprechenden Erwartungen, weshalb die gesamte Qualitätsmessung letztendlich ausschließlich auf die subjektiven Bewertungen fokussiert werden muss.

[3] OPERA™ … **O**bjective **Per**ceptual **A**nalyzer
http://www.opticom.de

3.1 Einleitung

Bei der digitalen Übertragung und Speicherung von Audiosignalen werden in zunehmendem Maße Datenreduktionsverfahren verwendet, die Eigenschaften des menschlichen Gehörs ausnutzen. Dabei wird versucht, die spektrale Verteilung der entstehenden Quantisierungsfehler so zu beeinflussen, dass sie unterhalb der Hörschwelle liegen. Die auf diese Weise unhörbar gemachten Störungen sind jedoch immer noch physikalisch vorhanden. Die wahrgenommene Qualität solcher gehörangepassten Codierverfahren kann somit mit konventionellen Messverfahren, die lediglich die insgesamt vorhandenen Störungen erfassen, nicht bestimmt werden. Daher wird die Qualität von gehörangepassten Codierverfahren üblicherweise mittels subjektiver Hörtests bestimmt. Solche Hörtests müssen unter optimalen Abhörbedingungen und mit einer großen Anzahl von Testhörern durchgeführt werden, so dass dieser Weg der Qualitätsbestimmung in vielen Fällen zu aufwendig ist. Ein objektives Messverfahren, das die zum subjektiven Qualitätseindruck führenden physiologischen und kognitiven Vorgänge simuliert, kann in vielen Fällen Abhilfe schaffen. Da hierzu verschiedene Vorschläge existierten, wurde in der ITU-R eine Arbeitsgruppe gegründet, die zum Ziel hatte, diese Vorschläge zu untersuchen und eine Empfehlung für ein Messverfahren zur „objektiven Messung der wahrgenommenen Audioqualität“ zu erarbeiten. Das für diese Empfehlung vorgesehene Messverfahren enthält sowohl Teile von zuvor existierenden Messverfahren, als auch eine Reihe von neuen Modellierungsansätzen.

Den grundsätzlichen Aufbau des Messverfahrens zeigt Abb. 3.1. Die Qualität des zu bewertenden Testsignals wird anhand seiner Abweichungen von dem als Referenz dienenden Originalsignal bestimmt. Dazu werden beide Signale in eine gehörangepasste Darstellung umgeformt (peripheres Gehörmodell). Anschließend werden durch einen Vergleich im Zeit- und Frequenzbereich verschiedene Abstandsmaße bestimmt. Diese Ausgangswerte („model output values“ - MOVs) werden zu einer einzelnen Kenngröße („distortion index“ - DI) zusammengefasst, die sich auf einen Schätzwert für die subjektiv empfundene Audioqualität abbilden lässt. Entsprechend der Bezeichnung der aus Hörtests gewonnenen Qualitätsbewertung als „subjective difference grade“ (SDG) wird dieser Ausgangswert als „objective difference grade“ (ODG) bezeichnet.

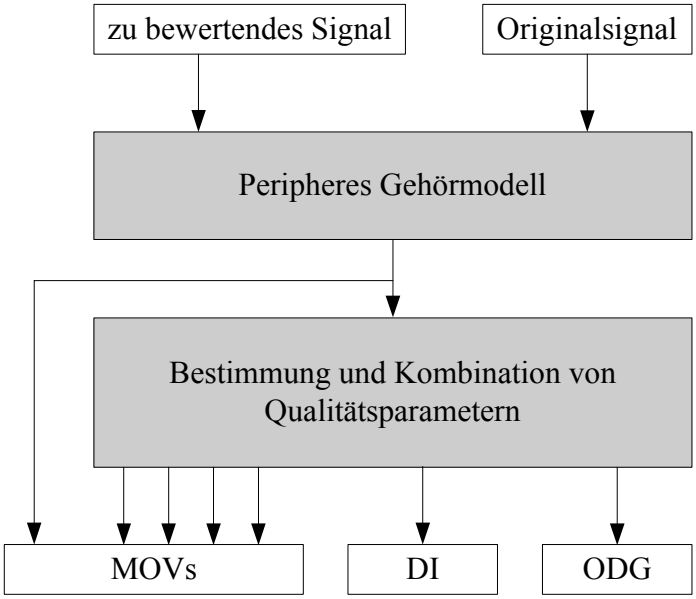

Abb. 3.1: Grundsätzlicher Aufbau des Messverfahrens

3.2 Arbeitsweise gehörangepasster Messverfahren

Es existieren zwei unterschiedliche Grundprinzipien, nach denen gehörangepasste Messverfahren arbeiten können: der Vergleich der Störung mit einer aus dem Originalsignal berechneten Maskierungsschwelle („masked threshold concept“) und der Vergleich zwischen gehörangepassten Signaldarstellungen von Originalsignal und zu bewertendem Signal („comparison of internal representations“). Als dritter Ansatz kann der direkte Vergleich der Spektraldarstellungen beider Signale (ohne Verwendung eines Gehörmodells) betrachtet werden.

3.2.1 Vergleich zwischen Maskierungsschwelle und Störung

Das Prinzip des Vergleichs der Störung mit einer Maskierungsschwelle („masked threshold concept“, auch: „noise signal evaluation“) wird in den ersten bekannten gehörangepassten Messverfahren verwendet. Dabei wird durch Subtraktion des Originalsignals vom zu bewertenden Signal das Fehlersignal berechnet und mit einer aus dem Originalsignal bestimmten Maskierungsschwelle verglichen. Vorteile dieses Konzeptes sind der relativ einfache Abgleich mittels aus psychoakustischen Experimenten gefundener Daten und die Verwendbarkeit des zugehörigen psychoakustischen Modells für Audiokodierverfahren.

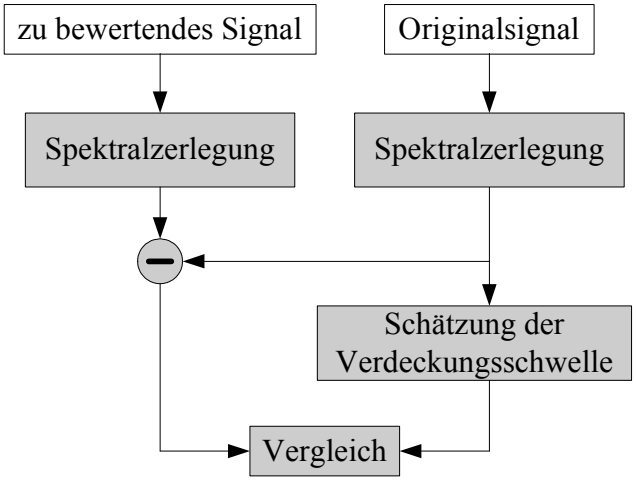

Abb. 3.2: Vergleich zwischen Maskierungsschwelle und Störung

3.2.2 Vergleich zwischen gehörangepassten Signaldarstellungen

Das Konzept des Vergleichs zwischen gehörangepassten Signaldarstellungen („comparison of internal representations“, auch: „comparison in the cochlear domain“), wird erstmals 1985 von Karjalainen [11] verwendet und bildet die Grundlage für die meisten neueren gehörangepassten Messverfahren. Dabei werden sowohl aus dem Originalsignal als auch aus dem zu bewertenden Signal gehörangepasste Darstellungen (so genannte Erregungsmuster) bestimmt. Die Bewertung der Qualität erfolgt aus dem Vergleich dieser Erregungsmuster. Diese Vorgehensweise kommt der physiologischen Arbeitsweise des Gehörs sehr viel näher als das zuvor beschriebene Konzept. Es bietet daher eine bessere Ausgangsbasis für komplexere Gehörmodelle.

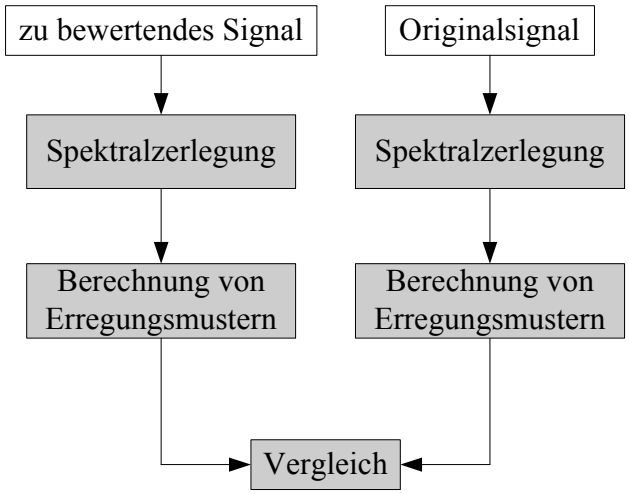

Abb. 3.3: Vergleich zwischen gehörangepassten Signaldarstellungen

3.2.3 Analyse von Fehlerspektren

Einige Effekte, wie z.B. die Wahrnehmung einer Grundfrequenz in Tonkomplexen, lassen sich anhand linearer Spektraldarstellungen einfacher modellieren als in einer dem Hörempfinden angepassten, nichtlinearen Frequenzskala. Ein solcher Ansatz kann zwar wegen des fehlenden Gehörmodells nicht die alleinige Grundlage eines Messverfahrens sein, er kann ein solches Modell aber ergänzen, da er zusätzliche Informationen über das Testsignal liefert, die aus einem Gehörmodell nur schwer gewonnen werden können.

Auf das periphere Gehörmodell und allen dazugehörigen Eigenschaften wird nicht weiter eingegangen, da dies nicht Themengebiet der Diplomarbeit ist. Für ausführlichere Angaben wird auf [08] und [09] verwiesen.

3.3 „Perceptual evaluation of audio quality” – PEAQ

Vorweg ein kurzer Überblick über die Vorläufer von PEAQ:

- Noise Loudness – NL (1979 Schroeder, Atal und Hall)
- Auditory Spectral Difference – ASD (1985 Matti Karjalainen)
- Noise-to-Mask Ratio – NMR (1987 Brandenburg)
- Perceptual Audio Quality Measure – PAQM
- Perceptual Evaluation – PERCEVAL
- Perceptual Objective Measure – POM
- Disturbance Index – DIX
- Objective Audio Signal Evaluation – OASE
- Toolbox

Das erste gehörangepasste Verfahren zur objektiven Bestimmung der Qualität von Audiosignalen wird Ende der 70er Jahre veröffentlicht. In den Folgejahren gibt es immer wieder Weiterentwicklungen, bis 1994 die ITU-R eine Arbeitsgruppe zum Thema „gehörangepasste Messverfahren“ gründet. Ein „call for proposals“ führt zu sieben Vorschlägen (DIX, NMR, OASE, PAQM, PERCEVAL, POM und Toolbox), die auf ihre Leistungsfähigkeiten überprüft werden sollen. In Folge der unbefriedigenden Ergebnisse des ITU-Vergleichstests wird beschlossen, die besten Elemente der verschiedenen Messverfahren in einer neuen Methode zu kombinieren.

Der PEAQ Standard entsteht in einer Zusammenarbeit aller an der Entwicklung der oben beschriebenen Messverfahren beteiligten Organisationen. Im peripheren Gehörmodell sind Teile aller genannten Verfahren enthalten; die Qualitätsparameter (vgl. [08], [09], [10]) stammen größtenteils aus DIX und NMR, aber auch aus PERCEVAL und OASE. Das neue Messverfahren enthält sowohl ein FFT-basiertes, als auch ein filterbankbasiertes Gehörmodell. Die Qualitätsparameter werden zum Teil aus berechneten Verdeckungsschwellen und zum Teil aus dem Vergleich zwischen gehörangepassten Signaldarstellungen gewonnen. Daneben sind auch Ausgangsparameter enthalten, die kein Gehörmodell verwenden, sondern auf einem direkten Vergleich von FFT-Spektren beruhen. Die einzelnen Qualitätsparameter werden mit Hilfe eines einfachen künstlichen neuronalen Netzes zu einer die globale Audioqualität beschreibenden Kenngröße zusammengefasst.

Um eine exakte Modellierung der Hörschwellen zu ermöglichen, wird aus dem Abhörpegel des Testsignals ein Skalierungsfaktor für die Eingangssignale berechnet. Falls der Abhörpegel nicht bekannt ist, wird ein Pegel von 92dB SPL für einen vollausgesteuerten Sinuston angenommen. Weiters muss sichergestellt werden, dass Testsignal und Referenzsignal keinen Zeitversatz aufweisen.
Es gibt zwei verschiedene PEAQ Modellvarianten: die so genannte „basic version“ ist für Anwendungen gedacht, die eine geringe Rechenzeit erfordern, und die „advanced version“ liefert die bestmögliche Schätzung der empfundenen Audioqualität auf Kosten eines deutlich erhöhten Rechenaufwands.

3.3.1 „Advanced version“

Die „advanced version“ von PEAQ benutzt das filterbankbasierte Gehörmodell für die Bestimmung aller Qualitätsparameter, die durch einen Vergleich gehörangepasster Signaldarstellungen gewonnen werden und das FFT-basierte Gehörmodell für die übrigen Qualitätsparameter. Zu der erstgenannten Gruppe gehören die partielle Lautheit additiver Störungen, die partielle Lautheit linearer Verzerrungen und das Maß für die Veränderung der zeitlichen Hüllkurven. Zur zweiten Gruppe gehören die „Noise-to-Mask Ratio“ und das Maß für die harmonische Fehlerstruktur (Definition der Parameter siehe [08]). Die in der „basic version“ verwendeten Maße für die Wahrscheinlichkeit und Häufigkeit von hörbaren Verzerrungen sind in der „advanced version“ nicht notwendig.

3.3.2 „Basic version“

Die „basic version“ von PEAQ benutzt ausschließlich das FFT-basierte Gehörmodell für die Bestimmung der Qualitätsparameter. Die wegen der gröberen zeitlichen Auflösung der FFT fehlende Detailinformation wird zum Teil durch Verwendung einer größeren Anzahl von Qualitätsparametern (vgl. 3.3.1) ausgeglichen.

3.4 OPERA™

Zur Bestimmung, welche Messmethode (PEAQ oder PESQ[4] bzw. PSQM[5]) angewandt werden soll, kann man sich zwei Fragen stellen:

- Wird ein breitbandiges Audiosignal (Bandbreite > 16kHz) bewertet?
- Kann der Proband das zu testende Signal mit dem Originalsignal vergleichen, d.h. ist für den Test das Originalsignal verfügbar?

Werden beide fragen mit „JA“ beantwortet, wird der PEAQ-Algorithmus verwendet (wie in unserem Fall). Das Diagramm in Abb. 3.4 zeigt den schematischen Aufbau eines Hörversuchs mit PEAQ bzw. die Auswahlkriterien für eine Anwendung dieses Algorithmus’.

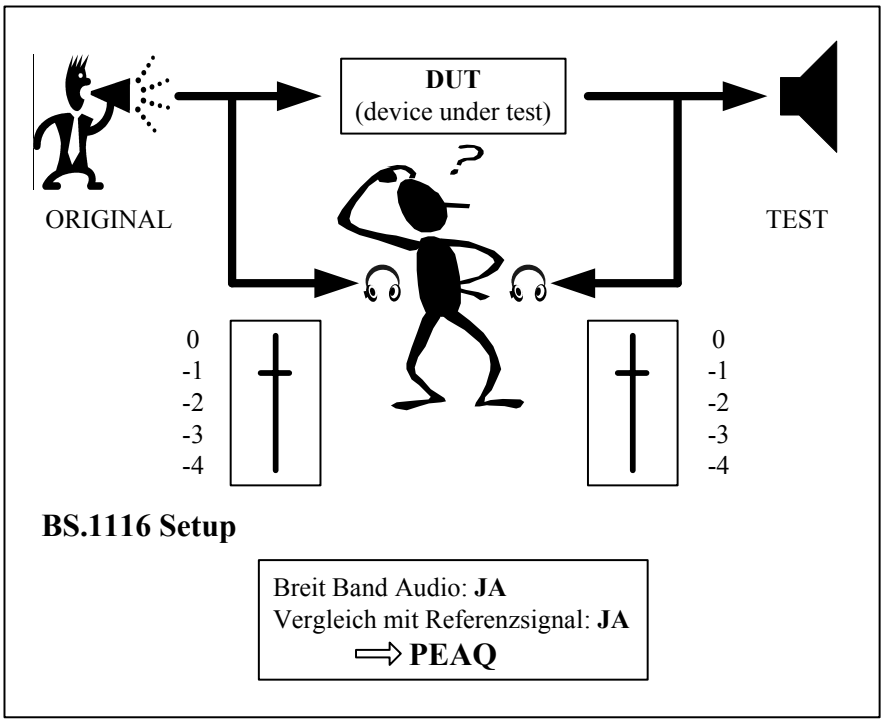

Abb. 3.4: Veranschaulichung der Arbeitsweise von BS.1116

[4] PESQ … „**p**erceptual **e**valuation of **s**peech **q**uality” (vgl. [10])
[5] PSQM .. „**p**erceptual **s**peech **q**uality **m**easure” (vgl. [10])

Die Standards ITU-T P.861 und ITU-R BS.1387 repräsentieren den neuesten Stand der Technik in Bezug auf objektive Bewertung von Audioqualität. Beide Techniken bzw. deren mathematische Modelle werden anhand entsprechender subjektiver Experimente ermittelt. Einen graphischen Überblick liefert Abb. 3.5.

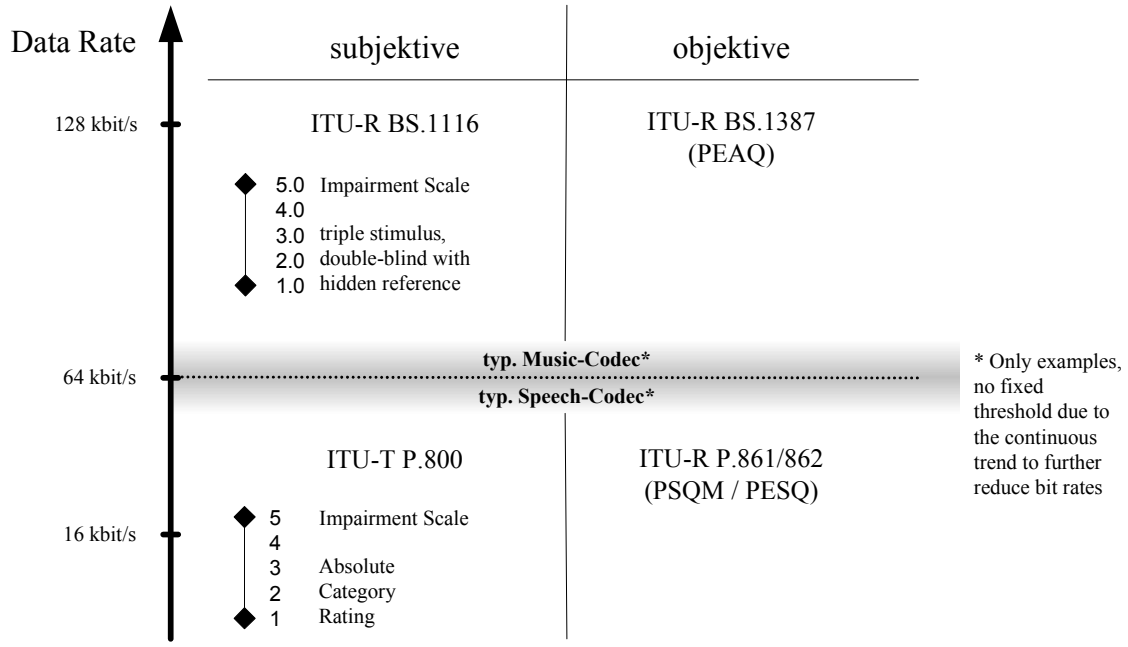

Abb. 3.5: Übersicht über subjektive und objektive Richtlinien

3.4.1 Anwendung der Software

Auf Details der Eingabeparameter der Software OPERA^{TM} wird nicht eingegangen. Bei Interesse wird auf [10] verwiesen.

Mit Hilfe der ermittelten SDGs (Hörversuch) der einzelnen Audiobeispiele (vgl. Tabelle 4.2, Tabelle 4.4) kann man ein Toleranzschema entwerfen und damit die berechneten ODGs (Software) auf ihre Genauigkeit überprüfen. Ziel wäre der Vergleich der subjektiven mit den objektiven Ergebnissen gewesen. Trotz intensiver Auseinandersetzung mit OPERA^{TM} ist dies jedoch nicht gelungen. Erst ein email an den Support der Firma konnte Klarheit schaffen:

Beispiel:

Es wird der ODG des Audiobeispiels AB1_1 (Signal mit unverschleierten DOs, vgl. Tabelle 4.3) mittels OPERA^{TM} („advanced version“) ermittelt. Man bekommt einen ODG(AV)=-0.45, obwohl insgesamt 17 DOs von 1-10 ms im Signal vorhanden sind! Ein Signal mit unverschleierten Ausfällen kann nicht mit „imperceptible“ bewertet werden; dieses wird im

subjektiven Test auch mit „very annoying“ bewertet. Laut OPERA™ schneidet dieses Signal im Punkt Qualität im Verhältnis wesentlich besser ab, als Signale, die restauriert worden sind. Ausfallsverschleierte Signale haben laut Software ODGs von -0.064 bis -0.253.

Der Grund für dieses Ergebnis: Da die Störungen nur sehr klein sind, das Signal ansonsten aber von hoher Qualität ist, werden diese kurzfristigen Störungen herausgemittelt. Sie sollten laut Auskunft des Firmen-Supports dennoch sichtbar werden, wenn man in OPERA™ den Diagramm Typ „Masked Threshold“ und „ODG vs. Time“ wählt. Im vorliegenden Fall ist dies aber nicht zielführend bzw. liefert er auch kein sinnvolles Ergebnis, da die Auflösung zu klein ist. Dies ist eine Einschränkung des Standards.
Restaurierte Signale liefern in der Regel etwas weiter gestreute Abweichungen von der Referenz als das gestörte, unbehandelte Signal, eine Eigenschaft, die sich wiederum nicht mit dem subjektiven Empfinden in Einklang bringen lässt.

Zusammenfassend kann man sagen, dass das OPERA™-Programm (bzw. der PEAQ-Standard) nicht bzw. nur sehr eingeschränkt für diese Problemstellung angewandt werden kann. Aufgrund dieser Fakten wird auf den Vergleich SDGs vs. ODGs verzichtet bzw. der Vergleich kann nicht durchgeführt werden.

(Bemerkung: Da die Untersuchung von Signalausfällen als Anwendungsbeispiel in der Software-Beschreibung genannt ist, wurde versucht, die genauen Spezifikationen geeigneter Signale zu erfahren bzw. direkt passende Testsignale zu bekommen. Beide Anliegen konnten von der Firma leider nicht zufrieden stellend erfüllt werden.)

4 Subjektiver Bewertungstest

Dieses Kapitel beschäftigt sich mit den Grundlagen bzw. Voraussetzungen, der Durchführung und der Auswertung des ersten subjektiven Bewertungstests, auch Hörversuch genannt. Zu Beginn wird auf die Richtlinien, die eine Grundlage des Versuchs sind, eingegangen. Nach deren Behandlung erfolgt eine genaue Dokumentation der Messungen und der daraus resultierenden Ergebnisse.
Die subjektiven Bewertungstests basieren auf den Richtlinien der Rec. ITU[6]-R BS.1116-1 [02]. Im Folgenden werden die wichtigsten Bestimmungen erläutert.

4.1 Grundlagen zum Hörversuch

Anhand von Hörversuchen ist es möglich, den Grad der Störung eines Audiosignals zu beurteilen. Beim Entwurf eines Bewertungstests muss darauf geachtet werden, dass die Bedienung und der Ablauf des Versuchs so einfach wie möglich realisiert werden. Überfordert man die Versuchsperson (VP), kann sich dies auf die Zuverlässigkeit ihrer Beurteilung auswirken.
Besteht ein Hörversuch aus Vergleichen zweier oder mehrerer Signale, ist eine weitere wichtige Bedingung für aussagekräftige Ergebnisse das Miteinbeziehen von so genannten Placebobeispielen, deren Art sich je nach Versuch unterscheidet. Typische Placebobeispiele sind entweder unbeeinträchtigte Audiosignale, sie entsprechen dem Referenzsignal, oder z.B. fehlerhafte Signale (gilt speziell für den vorliegenden Anwendungsfall), die noch nicht bearbeitet worden sind. Diese Überprüfungssignale werden zufällig miteinbezogen, so dass die VP keine Möglichkeit hat sie vorherzusagen.
Durch Vergleichen der Bewertungen der Placebobeispiele mit denen der tatsächlich beeinträchtigten Signale kann die Gültigkeit der Testergebnisse überprüft werden. Sollten zu

[6] ITU … **I**nternational **T**elecommunication **U**nit
http://www.itu.int/home/index.html

extreme Abweichungen auftreten, steht dem Versuchsleiter die Möglichkeit zur Verfügung, gewisse Datensätze zu vernachlässigen (siehe Abschnitt 4.1.1).

4.1.1 Auswahl der Versuchspersonen

Grundvoraussetzung eines jeden Versuchs ist die Auswahl geeigneter Versuchspersonen (VPN). Die Ergebnisse von Hörversuchen, die sehr geringe Beeinträchtigungen der Testsignale beurteilen, sollen ausschließlich von Probanden mit entsprechender Erfahrung gewonnen werden, sogenannten „expert listeners“. Diese VPN haben im Gegensatz zu „non-expert listeners“ Erfahrung im kritischen Hören und Bewerten von Audiosignalen. Dadurch bekommt man ein zuverlässiges und auch kritisches Ergebnis, als von ungeübten Probanden. Die Streuung der Daten von „expert listeners“ darf im Allgemeinen als geringer vorausgesetzt werden als die von „non-expert listeners“. Je höher die Qualität des zu testenden Systems ist, desto wichtiger wird der Einsatz von „expert listeners“.

In manchen Fällen ist eine Selektion der VPN notwendig. Diese Auswahl kann entweder vor („pre-screening“) oder nach („post-screening“) dem Test erfolgen. Es kann auch vorkommen, dass beide Verfahren angewandt werden.
Jede Anwendung muss davor gründlich überlegt werden, da einseitige und somit ungültige Testergebnisse die Folge von unsachgemäßen Aussonderungen sind. Die angewandten Auswahlkriterien werden in dem Testbericht klar ersichtlich beschrieben, um dem Leser einen Einblick zu ermöglichen.

4.1.1.1 Vorauswahl der VP („pre-screening“)

„Pre-screening“-Verfahren beinhalten Audiometrietests, die Auswahl der VP im Bezug auf ihre Erfahrung und Leistung in vorherigen Versuchen und die Auswahl einer VP anhand statistischer Analysen aus früheren Versuchen. Weiters kann auch die Einführungsphase des Versuchs zur Selektion herangezogen werden.
Das Hauptargument für eine Vorauswahl ist die damit verbundene steigende Effizienz von subjektiven Hörversuchen. Dies muss jedoch mit dem Risiko abgewogen werden, die Bedeutung des Ergebnisses dadurch einzuschränken bzw. im extremsten Fall zu verfälschen.

4.1.1.2 Auswahl der VP nach dem Versuch („post-screening“)

„Post-screening“-Verfahren werden in zwei Klassen unterteilt:

- Eine VP vergibt in regelmäßigen Abständen die gleichen Bewertungen.
- Es ergeben sich Widersprüche in der Bewertung einer VP, verglichen mit dem Mittelwert aller Probanden.

Sollten einige wenige VPN nur mit den Extrema der Skala bewerten, die Mehrzahl der Bewertungen ist jedoch um einen anderen Punkt konzentriert, so kann man diese als Ausreißer ansehen und die Daten verwerfen. Probanden, deren Beurteilungen sich nur am oberen Ende der Skala bewegen, werden eher als zu unkritisch eingestuft, während Probanden, die nur den unteren Skalenbereich nutzen, als zu kritisch betrachtet werden. Vernachlässigt man diese VPN, wird das Ergebnis der Auswertung realistischer.

Diese Methoden finden hauptsächlich dann Anwendung, wenn einzelne Probanden keine sachgemäße Beurteilung durchführen. „Post-screening“-Methoden können die Aussage des Testerergebnisses verändern. Man sollte sich deshalb die individuelle Empfindlichkeit der VPN gegenüber verschiedenen Artefakten im Gedächtnis behalten und dementsprechend Vorsicht bei der Selektion walten lassen.
Durch Erhöhen der Anzahl der zu testenden Personen werden die Auswirkungen einzelner extremer Bewertungen vermindert; daraus folgt eine Minimierung der Anwendung der Auswahlverfahren.

4.1.1.3 Anzahl der VPN

Prinzipiell werden „expert listeners“ immer den „non-expert listeners“ vorgezogen. Die Nichtexperten repräsentieren die breite Bevölkerungsmasse, während geschulte VPN übermäßig kritisch beurteilen. Eine wichtige Feststellung ist, dass viele „non-expert listeners“ im Laufe des Versuchs dazu neigen, sensibler diverse Artefakte zu beurteilen und so als Experten angesehen werden können. Daraus folgt, wie schon erwähnt, dass Versuche mit „expert listeners“ zu einem verbesserten Ergebnis führen.
Die Mindestanforderung an Probanden beträgt zwanzig „non-expert listeners“ und zehn „expert listeners“.

Allgemein gilt, dass vor jedem Versuchsbeginn eine Trainingsphase mit den Probanden durchgeführt wird. Dadurch gewöhnt sich die VP an den Testablauf, die Audiobeispiele und die Versuchsumgebung.

4.1.2 Wiedergabearten

Die Wiedergabe kann entweder über Lautsprecher (LS) oder über Kopfhörer erfolgen. Für beide Abspielarten gilt, dass der Grundgeräuschpegel im Raum so niedrig wie möglich gehalten wird und keine Störgeräusche (Trittschall, Lüfterrauschen, Sprache, etc.) auftreten.

4.1.2.1 Lautsprecher

Entscheidet man sich für diese Wiedergabeart, müssen alle ausschlaggebenden Informationen bezüglich der Dimension und der Nachhallzeit des Versuchsraumes, der Positionen der VPN im Raum und ihre Abstände von den LS angegeben werden.
Bei einer Bewertung von qualitativ hochwertigen Audiosignalen empfiehlt sich ein Wiedergabepegel von 80 bis 90 dB.

4.1.2.2 Kopfhörerwiedergabe

Bei der Kopfhörerwiedergabe kann man die Informationen über die Raumeigenschaften vernachlässigen. Wichtig sind aber auch hier genaue Angaben zu dem verwendeten Equipment.
Der Wiedergabepegel ergibt sich hier aus einem Vergleich von einem Referenzpegel (80 bis 90 dB) mit der wahrgenommenen Lautstärke im Kopfhörer. Die Wiedergabelautstärke im Kopfhörer soll der wahrgenommenen Referenzlautstärke entsprechen.

Die Angaben über die Testumgebung, die verwendeten Geräte und den Testablauf (siehe Abschnitt 4.1.3) dienen dazu, dass bei eventuell auftretenden Widersprüchen (Einsprüchen) der Versuch unter den gleichen Bedingungen reproduziert werden kann. Ausführlichere Informationen über die Wiedergabe mit LS oder Kopfhörern findet man in [01] bzw. [02].

4.1.3 Versuchsablauf

Zur Bewertung von sehr kurzen Störungen in Audiosignalen eignet sich die „double-blind triple-stimulus with hidden reference“ Methode besonders gut. Dieses Verfahren liefert eindeutige Ergebnisse und erlaubt, wie schon erwähnt, eine genaue Erkennung und Beurteilung von sehr kurzen Signalstörungen.

4.1.3.1 Versuchsbeschreibung

Der Bewertungstest wird von jeder VP einzeln durchgeführt. Es stehen drei Stimuli A, B und C zur Verfügung. Das Referenzsignal befindet sich immer an erster Stelle und wird mit A bezeichnet. Die Signale B und C stellen das Testsignal und das (versteckte) Originalsignal dar. Bei jedem neuen Beispiel werden B und C zufällig den beiden Signalen zugeteilt, sodass man keine Information über das zu evaluierende Beispiel hat. Die VP kann sich die einzelnen Signale beliebig oft anhören, um zwischen A und B, A und C und B und C Qualitätsunterschiede festzustellen.

Die Unterschiede zwischen A und B und zwischen A und C werden mittels der 5-teiligen ITU-R Bewertungsskalen (Abb. 4.1, [03]) eingestuft. Zwischenschritte sind im Zehntelbereich möglich. (Im Idealfall gibt es eine Bewertung mit Null und eine beliebige.)

Impairment	Grade	SDG[7]
Imperceptible	5.0	0.0
Perceptible, but not annoying	4.0	-1.0
Slightly annoying	3.0	-2.0
Annoying	2.0	-3.0
Very annoying	1.0	-4.0

Abb. 4.1: ITU-R five-grade impairment scale

Nachdem die VP ihre Bewertung eingegeben und bestätigt hat, erfolgt automatisch die Weiterleitung zum nächsten Beispiel, auch „Trial“ genannt. Der Versuch endet selbständig, nachdem alle Trials durchlaufen und bewertet wurden.

4.1.3.2 Eingewöhnungsphase / Trainingsphase

Vor dem eigentlichen Versuchsdurchlauf (VD) muss jede VP eine Trainingsphase absolvieren. Diese dient dazu, um mit der Versuchs-Hard- und Software, dem Versuchsablauf, dem Bewertungsprozess, der Skala und ihrer Bedeutung vertraut zu werden. Weiters werden der VP die verschiedenen Audiobeispiele mit den zu beurteilenden Artefakten vorgespielt. Während der Eingewöhnungsphase können auch noch auftretende Fragen beantwortet werden.

Auch Lerneffekte spielen eine wichtige Rolle bei der subjektiven Beurteilung von Audiosignalen. Eine unbekannte Störung wird deutlich schwerer erkannt als bekannte Störungsarten. Weiterhin zeigt sich beim Erkennen kleiner Störungen in sehr komplexen

[7] SDG … „**s**ubjective **d**ifference **g**rade”: $SDG=Grade_{signal\ under\ test}-Grade_{reference\ signal}$

Audiosignalen ein deutlicher Trainingseffekt. Je ausführlicher die Trainingsphase gestaltet wird, desto schneller können die VPN als „expert listeners“ betrachtet werden.

4.1.3.3 Testphase

Für genaue Ergebnisse wird der subjektive Hörversuch immer von einer VP alleine durchgeführt. Dadurch kann die VP ihr eigenes Tempo wählen, mit dem sie zwischen den Stimuli umschaltet. Eventuell auftretende Klick-Geräusche durch das Umschalten oder Bewerten sind zu vermeiden, da sie eine negative Auswirkung auf den Bewertungsprozess haben können.
Die Testphase sollte nicht länger als 20-30 Minuten dauern. Zu lange Versuche führen zur Ermüdung der VP und so zu verfälschten Ergebnissen.

4.2 Versuchsdurchführung

Dieses Unterkapitel beschäftigt sich mit der Durchführung des subjektiven Bewertungstest und seinen Modifikationen gegenüber der in 4.1 beschriebenen Richtlinien. Der Versuch ist, wie in Abb. 4.2 dargestellt, aufgebaut.

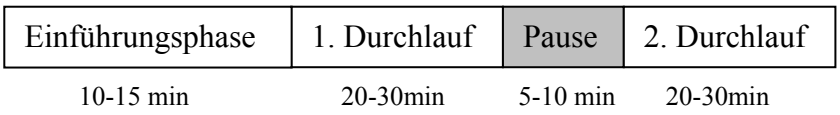

Abb. 4.2: Schematische Darstellung des Versuchsablaufs

Als Probanden werden großteils Toningenieur-Studierende ausgesucht. Diese können auf Grund des Studiums als kritische Hörer eingestuft und dadurch eher der Kategorie „expert listeners“ als „non-expert listeners“ zugeteilt werden. Der zweite Kreis beinhaltet Probanden, die sich täglich und vorwiegend im Beruf, mit Musik beschäftigen und/oder ein Instrument spielen. Die letzte und kleinste Gruppe besteht aus Personen, die sich weder mit Musik beschäftigen, noch ein Instrument spielen. Insgesamt ergeben sich somit 40 Probanden.

Toningenieur-Studierende	25 (1)
Personen mit „musikalischer Erfahrung“	04 (1)
Personen ohne „musikalische Erfahrung“	11 (2)
Probanden	**40 (4)**

Die Zahlen in Klammer sagen aus, wie viele davon Frauen sind. Der Anteil an weiblichen VPN beträgt 10%. Von den vierzig Probanden sind fünf Probanden über dreißig Jahre alt, das entspricht 12.5%.
Durchgeführt wird der subjektive Bewertungstest im Experimentalstudio des IEM. Der gemessene Ruheschallpegel des Studios liegt bei 34.1dB$_{LAeq}$[8] (die Messdauer beträgt drei Minuten). Als Wiedergabeart wird die Kopfhörerwiedergabe verwendet. Damit kann man wesentlich genauer sehr kleine Signalunterschiede feststellen. Eine Dämpfung von zufällig auftretenden Störgeräuschen (Knackser im Raum durch die Belüftungsanlage) ist mit den verwendeten offenen Kopfhörern nicht möglich. Diese Geräusche haben jedoch keinen Einfluss auf die Ergebnisse, da sie klar unterscheidbar von den zu beurteilenden Artefakten sind.

Eingestellte Parameter des Pegelmessers:
Time Weighting Slow (long attack and release time)
Range: 20-100dB$_{SPL}$

Messdauer: Clock 00:03:00
Gemessener SPL = 34.1dB$_{LAeq}$

Der Versuch wird in zwei Versuchsserien (VSN) mit jeweils zwanzig Probanden durchgeführt. Unterschiede in den Serien gibt es beim Einführungstest und bei der Bedienung des Computers. Detailiertere Angaben kann man Abschnitt 4.2.4 entnehmen.

4.2.1 Messung der Kopfhörerdämpfung am Kunstkopf

Da im Datenblatt des Kopfhörers keine Angaben über den Dämpfungsgrad vorhanden sind, wird dieser an Hand einer Messung am Kunstkopf ermittelt. Mit Hilfe des Programms Cool Edit Pro 2.0 wird Rosa Rauschen (Intensity 12) erzeugt. Der Pegel am Kunstkopf-Verstärker wird auf 0dB eingestellt (Kunstkopf ohne aufgesetzte Kopfhörer). Aus der Pegeldifferenz, mit und ohne Kopfhörer, kann man nun die Dämpfung berechnen.
Der Pegelunterschied ist jedoch so klein, dass er auf der Anzeige des Kunstkopfverstärkers nicht ersichtlich ist. Deshalb kann die Dämpfung des Kopfhörers vernachlässigt werden.

[8] dB$_{LAeg}$ … „equivalent continuous SPL“ oder „time-average sound level“

Verwendete Geräte:

- MINILYZER ML 1, Inv.Nr.: Kunstuniversität Graz 540-1/2/4-01
- MiniSPL, Inv.Nr.: Kunstuniversität Graz 574-4/15/8-02
- Electrostatic Audio Products STAX[9], SRM-007t
  VACUUM TUBE OUTPUT DRIVER UNIT
  Inv.Nr.: Kunstuniversität Graz 574-4/15/26-01
- Electrostatic Audio Products STAX, SR-007 (OMEGA II)
  Electrostatic Earspeaker
  Inv.Nr.: Kunstuniversität Graz 574-4/15/26-01
- Laptop: Sony Vaio PCG-FX800, Mobile AMD Athlon™, XP 1600+
  1.40GHz, 256 MB RAM
  Betriebssystem: MS Windows XP Home, SP2
  MATLAB® Version 6.0.0.88 Release 12

4.2.2 Audiobeispiele

Die Audiobeispiele werden so ausgewählt, dass damit ein möglichst großer Musikbereich (Anwendungsbereich) abgedeckt ist. Weitere Kriterien, die bei der Auswahl zu berücksichtigen sind, sind obertonreiche Instrumente, Soloinstrumente, Jazzcombo mit Soloinstrument und „viel high head“, Orchester und Sprache.

Ausgewählte Audiobeispiele:

AB1.wav … Soundtrack Kill Bill Vol.2 – „Goodnight Moon” – by Shivaree

AB2.wav … Blackbird – „Grey – The World” – by Heinrich von Kalnein

AB3.wav … Ludwig van Beethoven, Symphonie Nr.5 c-moll op. 67, Ouvertüre „Leonore II”, Wiener Philharmoniker – „Allegro”

AB4.wav … EBU[10]: tec_sqam_39a_bwf_tcm6-12544 (Piano)

AB5.wav … EBU: tec_sqam_08m_bwf_tcm6-12488 (Violine)

AB6.wav … EBU: tec_sqam_53_bwf_tcm6-12476 (Sprache, weiblich, deutsch)

Bei den MATLAB®-Diagrammen in Abschnitt 4.3 werden die Audiobeispiele mit Pop, Jazz, Orchester, Piano, Violine und Sprache bezeichnet.

[9] http://www.stax.co.jp/
[10] European Broadcasting Union, Sound Quality Assessment Material (SQAM) Recordings For Subjective Tests http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/index.php

Von jedem Audiobeispiel wird ein passender fünf Sekunden langer Ausschnitt ausgewählt und in Cool Edit Pro 2.0 auf -9dB normalisiert. Um eine eventuell auftretende Signalverzögerung bei der Wiedergabe mit MATLAB® zu verdecken, werden alle Signale im Vorhinein um 50ms verzögert. Das Ein- bzw. Ausblenden der Signale liegt im Bereich von 200ms.

4.2.3 Algorithmen

Der durchgeführte subjektive Bewertungstest dient zur Evaluierung von zwei „Dropout-Concealment“-Algorithmen, die aktuell am IEM entwickelt werden. Zur Beschreibung der Funktionsweise der Algorithmen wird auf die Abschnitte 2.1.1.1 und 2.1.1.2 verwiesen.

4.2.3.1 Fehlerverteilung / Fehlerszenarios

Den beiden Algorithmen muss neben einigen anderen Parametern auch die Information übermittelt werden, ob es sich um eine gültige Signalinformation oder einen Ausfall handelt. Für die Simulation der Algorithmen wird angenommen, dass 2% des Gesamtsignals (vgl. Abschnitt 4.2.2) fehlerbehaftet sind. Mit diesem Prozentsatz und anhand der Fehlerverteilung (FV) aus Tabelle 4.1 werden vier Fehlerszenarios (FSs) berechnet.

Fehlerverteilung	
Länge in [ms]	Anzahl
1	2
2	2
3	1
4	2
5	2
6	1
7	2
8	2
9	1
10	2

Tabelle 4.1: VS1/VS2, Fehlerverteilung

Diese Fehlerszenarios sind normalverteilte Pseudozufallsfolgen, d.h. FS1 hat auf AB1 eine völlig andere Auswirkung als z.B. FS1 auf AB2. Bei der Auswertung der Hörversuche sieht man deutliche Unterschiede zwischen den einzelnen FSS (vgl. Abschnitt 4.3).

Für den subjektiven Bewertungstest ergeben sich somit 48 Trials (für eine genaue Auflistung der einzelnen Parameter vgl. Tabelle 4.2).

$$2\ \textit{Algorithmen} \times 4\ \textit{Fehlerverteilungen} \times 6\ \textit{Audiobeispiele} = \underline{\underline{48\ \textit{Trials}}}$$

4.2.4 Versuchsserien (VSN)

Bevor der eigentliche Versuch entworfen wird, werden fünf Vortests durchgeführt. Anhand der gewonnenen Informationen ist es möglich, die graphische Oberfläche, die Bedienung und den Testablauf bestmöglich zu gestalten.
Der Versuch kann in zwei VSN mit je zwanzig Probanden geteilt werden. Bei der ersten Serie (vgl. Abb. 4.4) kann die VP nur mittels Maus das Programm bedienen, während bei der zweiten (vgl. Abb. 4.6) auch die Tastatur miteinbezogen wird[11].
Nach Bestätigung des Testbeginns erscheint ein Play-Button („Play A-B / A-C“) und zwei Bewertungsskalen. Durch Betätigen dieses Play-Buttons (per Mausklick bzw. durch Drücken der Taste „a“) wird folgende Reihenfolge abgespielt:

Signal A	Pause	Signal B	Pause	Signal A	Pause	Signal C
5.05sec	0.5sec	5.05sec	1sec	5.05sec	0.5sec	5.05sec

Abb. 4.3: Wiedergabereihenfolge der Audiobeispiele

Das Referenzsignal befindet sich immer an erster Stelle und wird mit A bezeichnet. Die Signale B und C stellen das Testsignal und das (versteckte) Originalsignal dar. Bei jedem neuen Beispiel werden B und C zufällig den beiden Signalen zugeteilt, sodass man keine Information über das wahre Beispiel hat. Der Play-Button kann nur einmal betätigt werden. Während der Wiedergabe wird das jeweilige Bezeichnungsfeld des aktuellen Beispiels rot hinterlegt (siehe Abb. 4.5; das Referenzsignal wird wiedergegeben). Die Bewertung kann entweder während oder nach der Wiedergabe der Samples erfolgen. Erkannte Qualitätsunterschiede zwischen A und B und zwischen A und C werden anhand der 5-teiligen Bewertungsskalen (Abb. 4.1) eingestuft. (Im Idealfall gibt es eine Bewertung mit Null und eine davon verschiedene.)

[11] Das Miteinbeziehen der Tastatur als Bedienelement wird auf Anregung mehrerer VPN aus VS1 berücksichtigt.

Nach erfolgter Bewertung ermöglicht es der Next-Button (Mausklick bzw. Taste „d“), zum nächsten Trial zu wechseln. Insgesamt werden 48 bzw. 52 Signale evaluiert.

Diverse selbsterklärende Fehlermeldungen und Hinweise wie „Sind Sie bereit für den Test / Einführungstest?“, „Falsche Taste! Bitte Taste „a“ für Play benutzen!“, „Maximale Wiedergabeanzahl erreicht!“ und „Zuerst aktuellen Test durchführen!“ bieten der VP zusätzliche Informationen und Hilfe.

4.2.4.1 Erste Versuchsserie (VS1)

4.2.4.1.1 Einführungsphase

Bei der Einführungsphase der VS1 gibt es einen Bewertungsreferenzwert (siehe Abb. 4.4 rechts oben, rote Schrift). Der ursprüngliche Gedanke ist, die Werte aus den objektiven Bewertungen (Evaluierung mittels Computerprogramm) als Referenzwerte zu verwenden. Da die so ermittelten Ergebnisse (vgl. Abschnitt 3.4) nicht mit den Erwartungen übereinstimmen, werden die Referenzwerte entsprechend einer möglichen subjektiven Bewertung festgelegt. Dieser Wert stellt einen Richtwert für die Bewertung des aktuellen Trials dar und ermöglicht es, die VP auf die „ITU-R five-grade impairment scale“ grob zu kalibrieren.

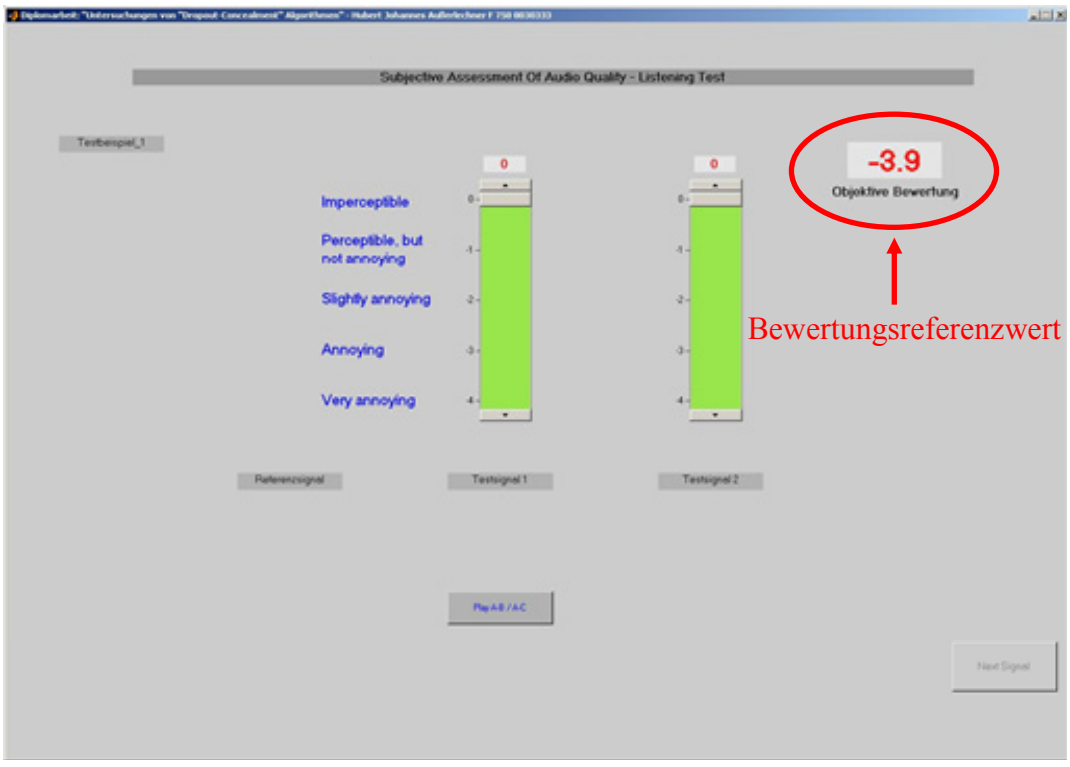

Abb. 4.4: Graphische Oberfläche des Einführungstests (Version 1)

Als Testbeispiele werden die sechs Audiobeispiele (vgl. Abschnitt 4.2.2) mit je einem unterschiedlichen FS implementiert. Die Reihenfolge wird so gewählt, dass die Qualität der berechneten Signale mit „very annoying“ aufsteigend bis „imperceptible“ zu beurteilen ist. Das erste Beispiel ist somit der schlechteste Fall (Signal mit DOs)!, vgl. Tabelle 4.3 die ersten sechs Beispiele), das letzte Beispiel der beste vorkommende Fall (Originalsignal). Der VP wird auch mitgeteilt, dass diese beiden Fälle im Haupttest nicht vorkommen (d.h. in VS1 sind keine Placebobeispiele enthalten)!
Ist ein Button dunkelgrau hinterlegt, kann er betätigt werden. Bei hellrauer Hintergrundfarbe muss zuerst der entsprechende Schritt durchgeführt werden.

Beispiel: Das erste Trial wird geladen. Der Play-Button ist dunkelgrau, der Next-Button hellgrau hinterlegt. Betätigt man den Next-Button, erscheint eine Aufforderung, „Zuerst aktuellen Test durchführen!“. Eine Bewertung der Beispiele vor Beginn der Wiedergabe ist auch nicht möglich. Nach der Wiedergabe werden der Play-Button hellgrau und der Next-Button dunkelgrau hinterlegt. Dies ist deshalb nötig, da eine VP auch zweimal mit Null bewerten kann. Die Bestätigung der eingegebenen Bewertung erfolgt gleichzeitig bei Betätigen des Next-Buttons.

4.2.4.1.2 Testphase

Anschließend an den Einführungstest und nach Klärung letzter Fragen beginnt die Testphase. Eine genaue Auflistung der verwendeten Audiobeispiele kann man Tabelle 4.2 entnehmen. Aus dem Ordner mit diesen Beispielen werden durch das MATLAB®-Programm zufällig die Audiofiles ausgewählt und wiedergegeben. Für programmtechnische Details und genaue Beschreibungen wird auf die Programme „hja_sasq_GUI5“ und „hja_sasq_GUI_main_program5“ verwiesen (siehe beigelegte CD).

SASQ – Testbeispiele							
Trial	Audiobeispiel	Fehlerszenario	Algorithmus	Trial	Audiobeispiel	Fehlerszenario	Algorithmus
1	1	1	1	25	1	3	1
2	2	1	1	26	2	3	1
3	3	1	1	27	3	3	1
4	4	1	1	28	4	3	1
5	5	1	1	29	5	3	1
6	6	1	1	30	6	3	1
7	1	1	2	31	1	3	2
8	2	1	2	32	2	3	2
9	3	1	2	33	3	3	2
10	4	1	2	34	4	3	2
11	5	1	2	35	5	3	2
12	6	1	2	36	6	3	2
13	1	2	1	37	1	4	1
14	2	2	1	38	2	4	1
15	3	2	1	39	3	4	1
16	4	2	1	40	4	4	1
17	5	2	1	41	5	4	1
18	6	2	1	42	6	4	1
19	1	2	2	43	1	4	2
20	2	2	2	44	2	4	2
21	3	2	2	45	3	4	2
22	4	2	2	46	4	4	2
23	5	2	2	47	5	4	2
24	6	2	2	48	6	4	2

Tabelle 4.2: VS1/VS2, Testübersicht

Die einzigen Unterschiede zum Einführungstest sind die Anzahl der Trials (48 statt 6) und die fehlende Referenzanzeige (vgl. Abb. 4.4 mit Abb. 4.5). Nach dem ersten Testdurchlauf gibt es 5-10 Minuten Pause, um ausgerastet in den zweiten Durchlauf zu starten. Die verwendeten Audiobeispiele sind bei beiden Durchläufen die gleichen. Da ihre Auswahl auf einem Zufallsprozess beruht, ändert sich nur die Reihenfolge der Beispiele. Die Reihenfolge ist zwar zufällig, trotzdem entsprechen 50% der Signale B dem Original- bzw. Testsignal, ebenso umgekehrt, 50% der Signale C entsprechen dem Original- bzw. Testsignal. Damit hat die VP eine 50%ige Ratewahrscheinlichkeit, wenn sie immer gleich bewertet.

Bei der anschließenden Auswertung kann man somit u.a. herausfinden, ob eine VP bemüht ist, die Qualität zu beurteilen, oder ob sie nur willkürlich bewertet

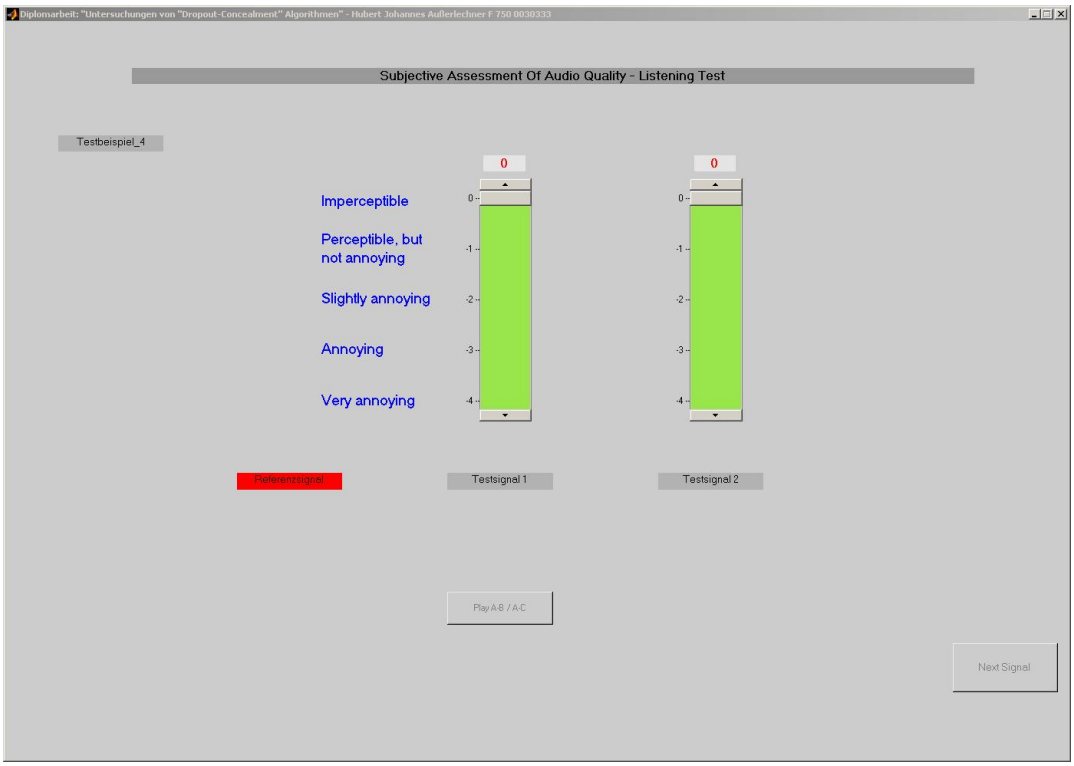

Abb. 4.5: Graphische Oberfläche des subjektiven Bewertungstests (Version 1)

4.2.4.2 Zweite Versuchsserie (VS2)

Die VS2 unterscheidet sich in drei Dingen von der ersten:

- Die Bedienung des Programms erfolgt mit Tastatur und Maus.
- Zusätzlich zu den berechneten Audiosignalen (entsprechen denen der VS1) werden vier Placebobeispiele eingefügt. Insgesamt ergeben sich somit 52 Beispiele. Die Placebobeispiele sind jeweils zwei Originalsignale und zwei unbehandelte Signale (Signale mit wirklichen DOs). Eine korrekte Bewertung würde beim ersten Fall 0 ergeben und beim zweiten Fall -4. Damit überprüft man die Aufmerksamkeit und die Empfindlichkeit des Probanden.
- Aufgrund von Kritik seitens der VPN der VS1 wird die Größe der Schriften auf der Oberfläche verändert (vgl. Abb. 4.5 mit Abb. 4.6).

4.2.4.2.1 Einführungsphase

Folgende Unterschiede ergeben sich zum Einführungstest der VS1:

- Wie aus Tabelle 4.3 ersichtlich, sind für diese Version des Einführungstests 12 Audiobeispiele vorgesehen.
- Die graphische Oberfläche (vgl. Abb. 4.6) des Einführungstests entspricht der des Haupttests (Wegfall des Bewertungsreferenzwertes).

Tabelle 4.3 zeigt die Unterschiede zu den Audiobeispielen der VS1. Hier wird der VP mitgeteilt, dass das erste Versuchsbeispiel dem schlechtesten vorkommenden Fall entspricht (Trial 1 und Trial 10: unberechnete, mit Ausfällen behaftete Audiobeispiele). Da in diesem Beispiel unverschleierte Signalausfälle vorhanden sind, ist anzunehmen, dass die Bewertung mit -4 erfolgt. Als zweites Placebobeispiel wird das Originalsignal eingefügt. Bei korrekter Bewertung wird der Wert 0 erwartet.

SASQ – Einführungstest Beispiele						
Trial	Audiobeispiel	Fehlerszenario	Algorithmus	Trial SASQ		
1	1	2	-	AB3_DO	mit DOs	AB1_1
2	2	4	2	AB4_46	-	AB2_2
3	3	4	1	AB6_42	-	AB3_3
4	4	3	1	AB2_26	-	AB4_4
5	5	4	1	AB5_41	-	AB5_5
6	6	-	-	AB1	Originalsignal	AB6_6
7	1	1	2	AB3_9	-	AB1_7
8	2	-	-	AB4	Originalsignal	AB2_8
9	3	2	2	AB6_24	-	AB3_9
10	4	1	-	AB2_DO	mit DOs	AB4_10
11	5	3	1	AB5_29	-	AB5_11
12	6	1	2	AB1_7	-	AB6_12

Tabelle 4.3: VS2, Audiobeispiele der Einführungsphase (Version 2)

Die letzte Spalte der Tabelle 4.3 zeigt die verwendete Nummerierung der Audiobeispiele für den Einführungstest (vgl. beigelegte CD).

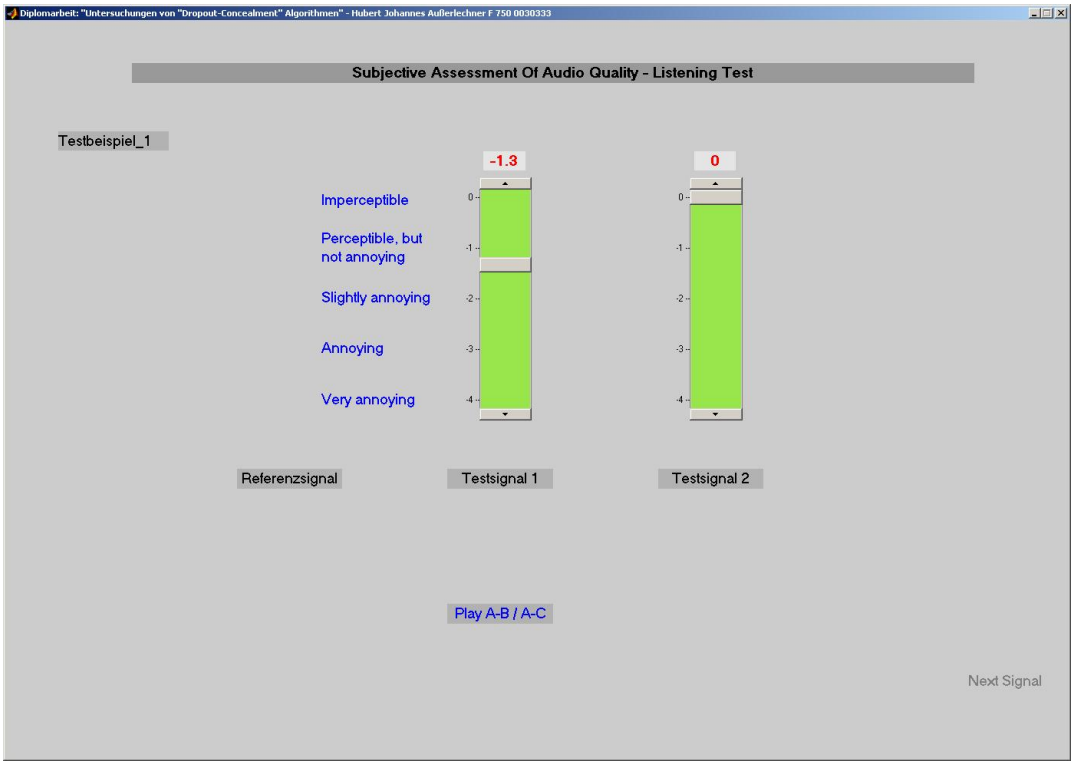

Abb. 4.6: Graphische Oberfläche des Einführungstests / subjektiven Bewertungstests (Version 2)

Zu Abb. 4.6: Diese Graphik zeigt eine Bewertung des Testsignals 1 mit -1.3 und des Testsignals 2 mit 0, d.h. die VP hat das zweite Beispiel als Originalsignal erkannt.

4.2.4.2.2 Testphase

Die Hauptunterschiede zur ersten Testserie wurden schon beschrieben. Tabelle 4.4 zeigt die verwendeten Placebobeispiele.

SASQ – Testbeispiele (Placebobeispiele)					
Trial	Audiobeispiel	Fehlerszenario	Algorithmus	Trial SASQ	
49	1	-	-	AB1	Originalsignal
50	4	-	-	AB4	Originalsignal
51	2	1	-	AB2_DO	mit DOs
52	3	2	-	AB3_DO	mit DOs

Tabelle 4.4: VS2, Placebobeispiele

4.3 Ergebnisse und Auswertung

Dieses Kapitel befasst sich mit der Auswertung der beiden VSN des subjektiven Bewertungstests. Zuerst wird die verwendete Auswertungsmethode erklärt.

4.3.1 Allgemeines zur Auswertung

Im Nachhinein gesehen ist eine verlängerte und somit intensivere Einführungsphase (siehe Einführungsphase des VD2) sinnvoller. Dadurch ist die VP mit dem Versuchsaufbau und –ablauf wesentlich besser vertraut. Der Lernprozess der VP fällt somit in die Einführungsphase und nicht in den Anfang der Versuchsphase. Durch die zufällige Auswahl der Trials pro VP und VD hat die VP zwar keine Möglichkeit, sich an die Reihenfolge der wiedergegebenen Audiobeispiele zu gewöhnen und im VD2 gleich zu bewerten, man kann dadurch aber auch einen eventuellen Lernprozess der VP nicht mehr untersuchen bzw. die ersten Trials in der Auswertung vernachlässigen.

4.3.1.1 Varianzanalyse

Die Varianzanalyse (**An**alysis **O**f **Va**riance, ANOVA) ist ein statistisches Verfahren der Datenanalyse und Mustererkennung, das versucht, die Varianz einer metrischen Variable durch eine oder mehrere Variablen zu erklären. Das Verfahren untersucht, ob (und gegebenenfalls wie) sich der Erwartungswert einer metrischen Zufallsvariable in verschiedenen Gruppen (auch Klassen) unterscheidet. In Prüfgrößen des Verfahrens wird getestet, ob die Varianz zwischen den Gruppen größer ist als die Varianz innerhalb der Gruppen. Dadurch kann ermittelt werden, ob die Gruppeneinteilung sinnvoll ist oder nicht bzw. ob sich die Gruppen signifikant unterscheiden oder nicht.

Voraussetzungen für eine Anwendung der ANOVA:

- Intervallskalen-Niveau der abhängigen Variable
- Normalverteilung
- Varianzhomogenität

Die ANOVA kann hier nicht angewandt werden da keine Intervallskala, keine Normalverteilung und keine Varianzhomogenität vorliegt. Die verwendete Messskala (0…-4) entspricht eher einer Ordinalskala und die Verteilung einer asymmetrischen Verteilung

(ungefähr einer um die Ordinate gespiegelten F-Verteilung). Weiters unterscheiden sich die Varianzen innerhalb einer Stichprobe[12] signifikant, d.h. Varianzhomogenität ist nicht gegeben.

Definitionen (vgl.[07]):

Nominalskala-„Ettiketierung“: Die Werte unterliegen keiner Rangfolge und sind nicht vergleichbar (Gleichheit, Verschiedenheit: z.B. Telefonnummer, Geschlecht, Rasse, Haltungsform).

Ordinalskala-„Ordnen“: Die Werte unterliegen einer Rangfolge, aber die Abstände zwischen den Werten der Skala lassen sich nicht bestimmen (Größer-kleiner-Relationen: z.B. Bewertung (Noten), Windstärke, Militärische Ränge).

Metrische Skala: Die Werte ((reelle) Zahlen) unterliegen einer Rangfolge und die Abstände zwischen den Werten der Skala lassen sich bestimmen (Gleichheit von Differenzen: z.B. Gewicht, Temperatur).

Intervallskala-„Beziffern“: Intervallskalen sind metrische Skalen, in denen über den Unterschied zweier Messwerte ausgesagt werden kann, ob er größer, gleich oder kleiner als der Unterschied zweier anderer Messwerte ist. Das bedeutet: Skalenwerte einer Intervallskala können bezüglich ihrer Differenzen (und Summen) verglichen werden.

4.3.1.2 Boxplot

Eine Möglichkeit, die Daten auszuwerten, ist das Programm SPSS. Da dieses Programm sehr komplex ist und auch mit MATLAB® die für diese Arbeit notwendigen Auswertungen möglich sind, werden die Messdaten mittels MATLAB® ausgewertet.

Verwendet wird der „boxplot“-Befehl (vgl. Abb. 4.7). Ein Boxplot (auch Box-Whisker-Plot) ist ein Diagramm, das zur graphischen Darstellung einer Reihe numerischer Daten verwendet wird. Es fasst verschiedene Maße der zentralen Tendenz, Streuung und Schiefe in einem Diagramm zusammen. Alle Werte der Fünf-Punkte-Zusammenfassung, also der Median, die zwei Quartile und die beiden Extremwerte, sind dargestellt.
Als „Box“ wird das durch die Quartile bestimmte Rechteck bezeichnet. Sie umfasst 50% der

[12] Eine Stichprobe stellt eine Untermenge einer Population (Gesamtmenge, z.B. die VPN) dar.

Daten. Durch die Länge der Box ist der Interquartilsabstand („interquartile range“, IQR) abzulesen. Dies ist ein Maß der Streuung, welches durch die Differenz des oberen und unteren Quartils bestimmt ist. Als weiteres Quartil ist der Median in der Box eingezeichnet, welcher durch seine Lage innerhalb der Box einen Eindruck von der Schiefe der den Daten zugrunde liegenden Verteilung vermittelt.
Mit „Whisker“ bezeichnet man die vertikalen Linien. Die Länge der Whisker beträgt maximal das 1,5-fache des Interquartilsabstands (Quartilabstand, $1.5 \cdot IQR$) und wird immer durch einen Wert aus den Daten bestimmt. Werte, die über dieser Grenze liegen, werden separat in das Diagramm eingetragen und als Ausreißer bezeichnet. Gibt es keine Werte außerhalb der Whisker, so wird die Länge des Whiskers durch den maximalen bzw. minimalen Wert festgelegt.
Häufig werden Ausreißer, die zwischen $1.5 \cdot IQR$ und $3 \cdot IQR$ liegen als „milde“ Ausreißer bezeichnet und Werte, die über $3 \cdot IQR$ liegen als „extreme“ Ausreißer. Diese werden dann auch unterschiedlich im Diagramm gekennzeichnet
Für alle folgenden Boxplot-Diagramme ist der Bezugspunkt mit Null definiert, d.h. die Bezeichnung „unterer/unterhalb“ bezieht sich auf jene Bereiche, die näher dem Wert Null liegen, während sich „oberer/oberhalb“ auf die Bereiche näher minus vier bezieht.

Die Robustheit gegenüber Datenausreißern ist ein weiterer Grund für die Anwendung des Boxplot-Befehls. Der Medianwert und der Quartilabstand lassen sich mit dem Mittelwert und der Varianz vergleichen. Der Mittelwert ist jedoch sensitiver gegenüber Ausreißern. Diese Eigenschaft erweist sich als Nachteil bei der Auswertung der Daten der Hörversuche.

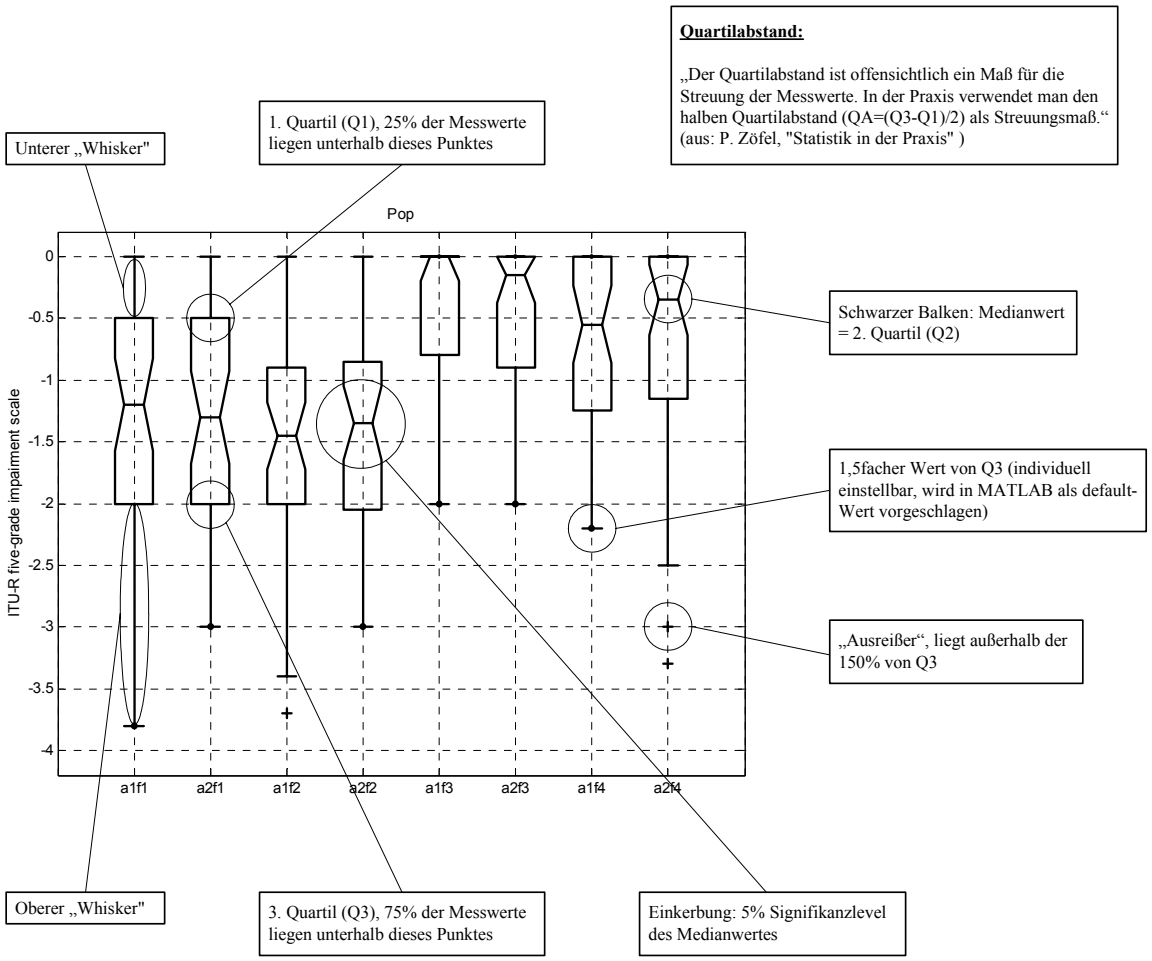

Abb. 4.7: Erklärung der Boxplot – Graphik

Unterer Grenzwert: $UG = Q1 - 1.5QA$ (4.1)

Oberer Grenzwert: $OG = Q3 + 1.5QA$ (4.2)

Quartilabstand = „interquartile range“: $QA = IQR$ (4.3)

Der Quartilabstand definiert die Größe des Bereichs, in dem die mittlere Hälfte der Daten liegen (zentraler 50%-Bereich) und eignet sich für eine Bewertung von ordinal skalierten Daten.

Vertrauensbereich, „confidence intervall“ (CI):

Der Medianwert $Q2$ liegt mit 95%iger Wahrscheinlichkeit innerhalb des CIs. Überschneiden sich die CIe zweier Datensätze, kann keine signifikante Aussage über die Differenz ihrer Medianwerte gemacht werden.

4.3.2 Auswertung der VS1

In den beiden folgenden Diagramme, Abb. 4.8 und Abb. 4.9, sind die Bewertungen des VD1 bzw. des VD2 der einzelnen Probanden abgebildet. Durch Vergleichen der beiden Graphiken zeigt sich eine Tendenz der schlechteren Bewertung des VD2 (eventuell aufgrund von Ermüdungserscheinungen). Vier bzw. fünf VPN (unter Berücksichtigung der minimalen Abweichung von $Q1_{VP\ Nr.18}$) weisen einen relative hohen Anteil an Fehlbewertungen (25%) im VD1 auf, da jeweils der UG mit dem Wert von $Q1$ zusammenfällt ($\mathrm{UG} = Q1 = 0$). Nur die Box von VP Nr.7 liegt bei beiden VDN am Wert Null. Keine Fehlbewertungen haben VP Nr.10 und Nr.11 im VD1 bzw. Nr.17 im VD2 (alle Probanden fallen in die Kategorie „expert listeners“).

Deutliche Unterschiede ergeben sich auch bei der Betrachtung der Skalenausnützung der einzelnen VPN. Während der Großteil die fünfteilige Skala relativ gut ausnutzt, erkennt man bei VP Nr.11 und Nr.19 im VD1 bzw. Nr.7 und Nr.10 (extreme Fokussierung) im VD2 einen eingeschränkten Bewertungsbereich.

Trotz einiger Verschlechterungen des VD2 werden für die weiteren Analysen die Daten beider Versuchsdurchläufe zusammengefasst, um eine größere Datenmenge und daraus resultierend ein gültiges Ergebnis zu bekommen.

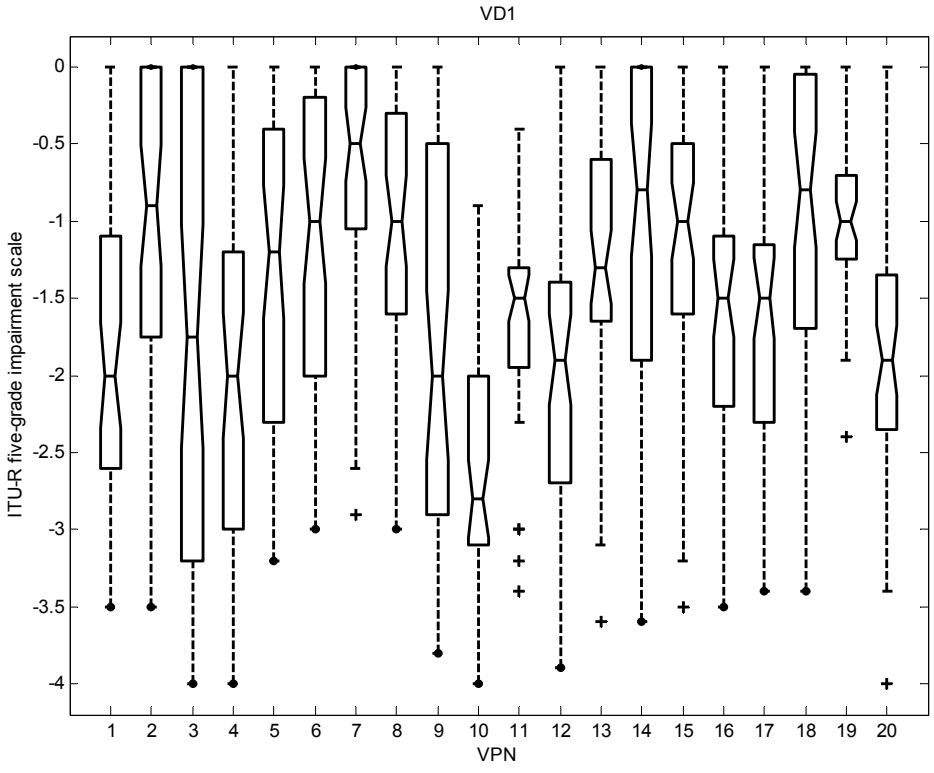

Abb. 4.8: VS1, Bewertung der einzelnen VPN, VD1

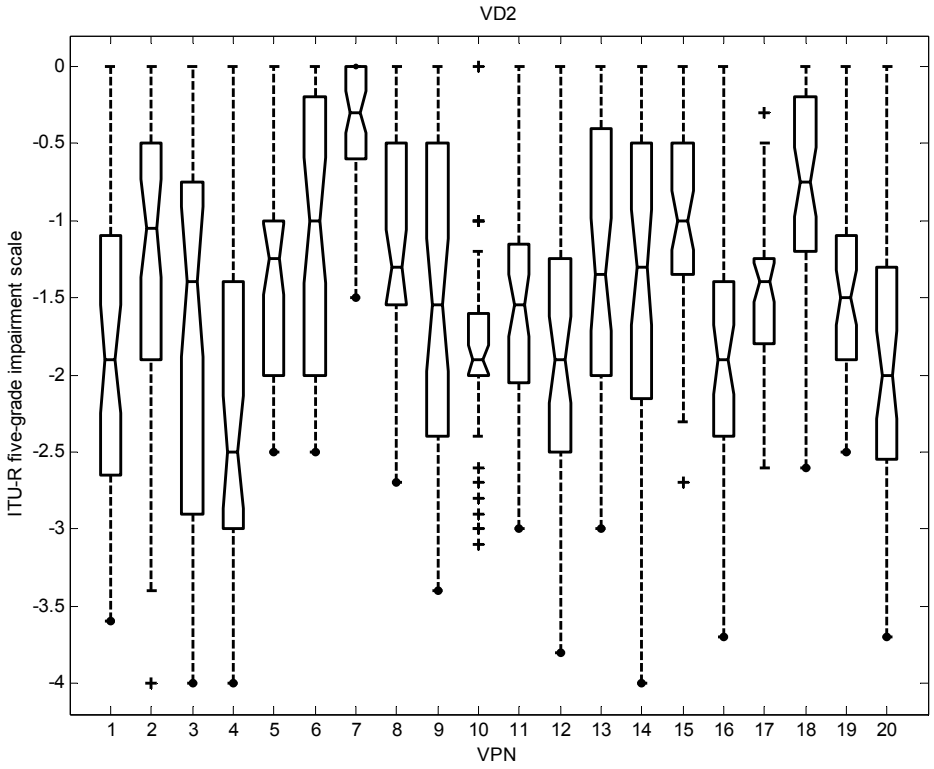

Abb. 4.9: VS1, Bewertung der einzelnen VPN, VD2

4.3.2.1 Auswirkung der FSs

Die erste Auswertung befasst sich mit der Auswirkung der vier FSs (siehe Abschnitt 4.2.3.1) auf die Bewertung der zwei Algorithmen und der Audiobeispiele. In Abb. 4.10 sind die Ergebnisse der zwei Algorithmen gegenübergestellt, wobei alle Trials eines FSs zusammengefasst sind. (Die Graphiken Abb. 7.1 und Abb. 7.2 zeigen die Ergebnisse des VD1 und des VD2 getrennt für die Algorithmen und für jedes Audiobeispiel über alle FSs.)

Man erkennt deutlich, dass jede Bewertungsverteilung den vollen Skalenbereich $[0,-4]$ ausnutzt. Bis auf minimale Unterschiede sind sich alle Boxplots ähnlich. Die Medianwerte pro FS weichen nur geringfügig voneinander ab; alle CIe überlappen sich (mit Ausnahme des CIs von a2f3) und liegen im Bereich $[-1.05,-1.85]$ („perceptible, but not annoying“, „slightly annoying“). Das Intervall der berechneten Medianwerte ist wesentlich geringer $[-1.2,-1.65]$ und deutet auch auf eine relativ gute allgemeine Bewertung der Audiobeispiele und im Weiteren der Algorithmen hin.

Verwendete Abkürzungen in den folgenden Graphiken:

axfy … „a“ bedeutet Algorithmus, „x“ steht für die entsprechende Nummer (1,2); „f“ bedeutet Fehlerszenario, „y“ entspricht wiederum der Nummer (1…4).

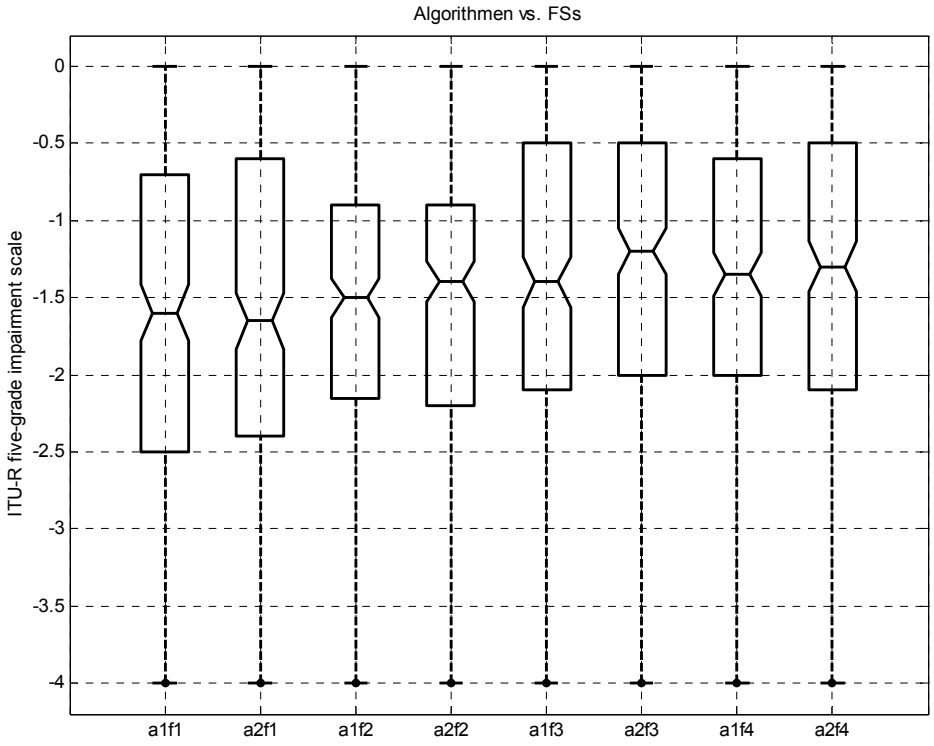

Abb. 4.10: VS1, Gegenüberstellung der zwei Algorithmen mit unterschiedlichen FSs

In den folgenden sechs Graphiken werden die Bewertungen der einzelnen Audiobeispiele dargestellt. Die Skalierung der Abszisse erfolgt mit Algorithmus eins bzw. zwei und den FSs (1-4).

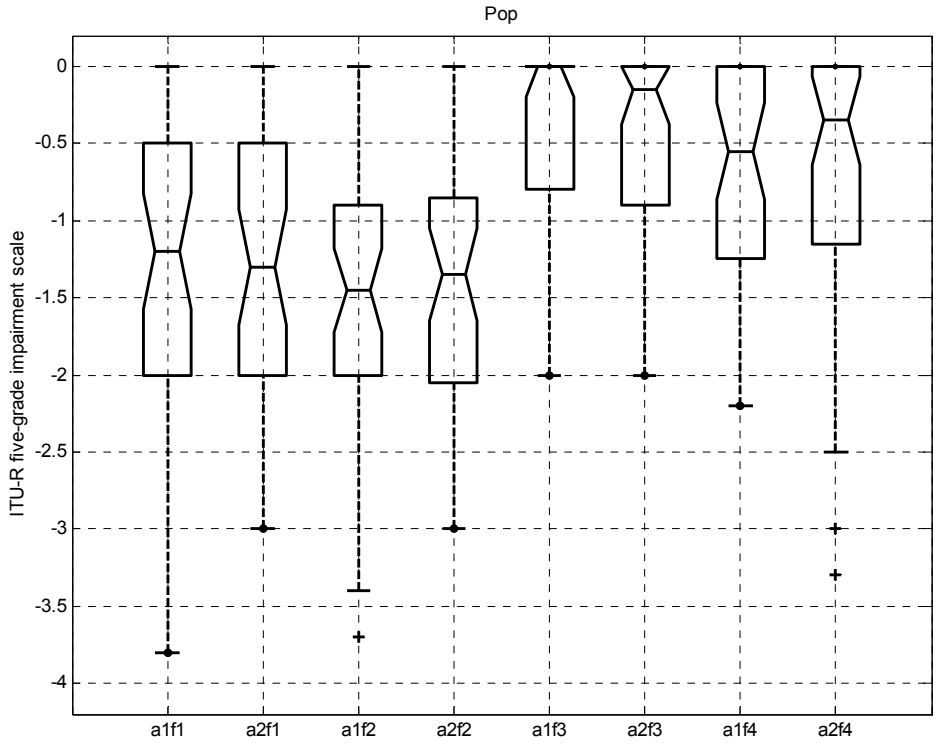

Abb. 4.11: VS1, Audiobeispiel Pop, Gegenüberstellung der Algorithmen (verschiedene FSs)

Man kann in Abb. 4.11 (Pop) deutlich zwei Gruppen erkennen: FS1 und FS2 weisen sowohl mit Alg1 als auch mit Alg2 eine negativere Bewertung auf, als FS3 und FS4. Auffallend ist bei Gruppe zwei, dass der UG und $Q1$ eines Boxplots übereinstimmen und den Wert Null haben („imperceptible“), d.h. in allen vier Fällen sind 25% der Bewertungen Fehlbewertungen. Einen noch höheren Prozentsatz liefern die Ergebnisse von a1f3 mit $\mathrm{UG} = Q1 = Q2 = 0$; dort werden 50% der verschleierten Signale nicht erkannt! Da bei a2f3 das CI mit dem Wert Null beginnt, ist es mit einer 95%igen Wahrscheinlichkeit möglich, dass hier auch $Q2_{a2f3} = 0$ ist.

Hört man sich die Audiobeispiele nochmals an, kann man einige verschleierte DOs bei den ersten zwei FSs deutlich erkennen. Diese befinden sich an Gesangsstellen und Stellen mit liegenden Akkorden bzw. Tönen. Bei den anderen zwei FSs fallen die DOs mehr oder weniger mit Schlagzeuggeräuschen zusammen und können dadurch schlechter wahrgenommen werden.

Die erste Gruppe nutzt den Skalenbereich für die Evaluierung aus, während die Bewertungen der zweiten Gruppe hautsächlich im Intervall $[0,-2]$ liegen. Aufgrund der genannten Tatsachen und da sich weder die Medianwerte, noch die CIe der beiden Gruppen überschneiden, folgt, dass das Popbeispiel mit FS3 und FS4 wesentlich besser bewertet wird,

als mit FS1 und FS2. Eine getrennte Beurteilung der Algorithmen kann aus dieser Graphik nicht erfolgen, d.h. beide Algorithmen können hier als gleichwertig angesehen werden.

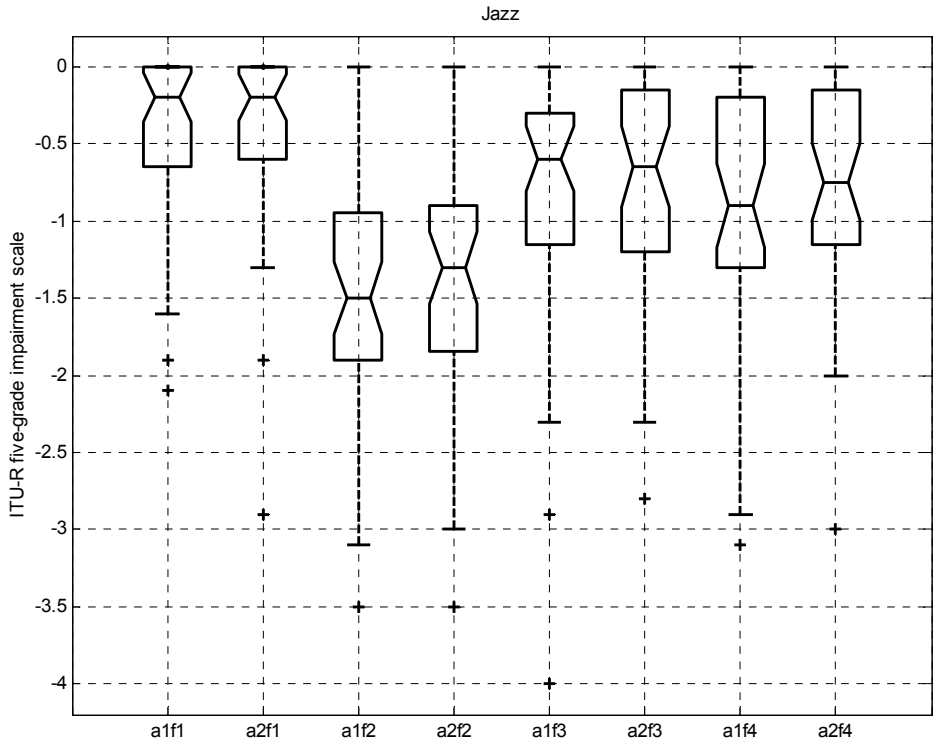

Abb. 4.12: VS1, Audiobeispiel Jazz, Gegenüberstellung der Algorithmen (verschiedene FSs)

Bei den ersten zwei Boxplots (FS1) der Bewertungen des Jazzbeispiels (vgl. Abb. 4.12) tritt wieder der Fall $UG = Q1 = 0$ auf (25% der Bewertungen sind Fehlbewertungen). Für die beiden Medianwerte gilt $Q2_{a1f1} = Q2_{a2f1} = -0.2$, d.h. 50% der Bewertungen liegen im Intervall $[0, -0.2]$. Berücksichtigt man noch die Werte von $Q3$, treten 75% aller Bewertungen innerhalb des Bereichs $[0, -0.6]$ („imperceptible“) auf. Die Skala wird hier bei Weitem nicht ausgenutzt.

Infolge der kleinen negativen Werte von $Q1$ des FS3 und FS4 lässt sich auf einen hohen Anteil an Fehlbewertungen bzw. auf viele Bewertungen im Intervall $[0, -0.3]$ schließen. Betrachtet man zusätzlich die Werte von $Q3$, können 75% der Bewertungen als „perceptible, but not annoying“ betrachtet werden. Die fehlerverschleierten Audiobeispiele des FS2 werden von allen vier FSs am schlechtesten beurteilt (die Boxen liegen aber trotzdem in der oberen Hälfte der Bewertungsskala).

Vergleicht man das Diagramm des Jazzbeispiels mit denen der anderen Audiobeispiele, bemerkt man, dass wesentlich mehr Ausreißer (11) vorhanden sind, als bei den anderen (0-3).

Ein Grund dafür wäre die schlechtere Wahrnehmbarkeit der verschleierten DOs infolge der starken Präsenz der Schlagzeuggeräusche (u.a. viel „high head“).
Im Allgemeinen ergibt sich eine sehr positive Bewertung des Jazzbeispiels. Aus den Ergebnissen geht hervor, dass die beiden Algorithmen auch bei diesem Beispiel gleichwertig arbeiten.

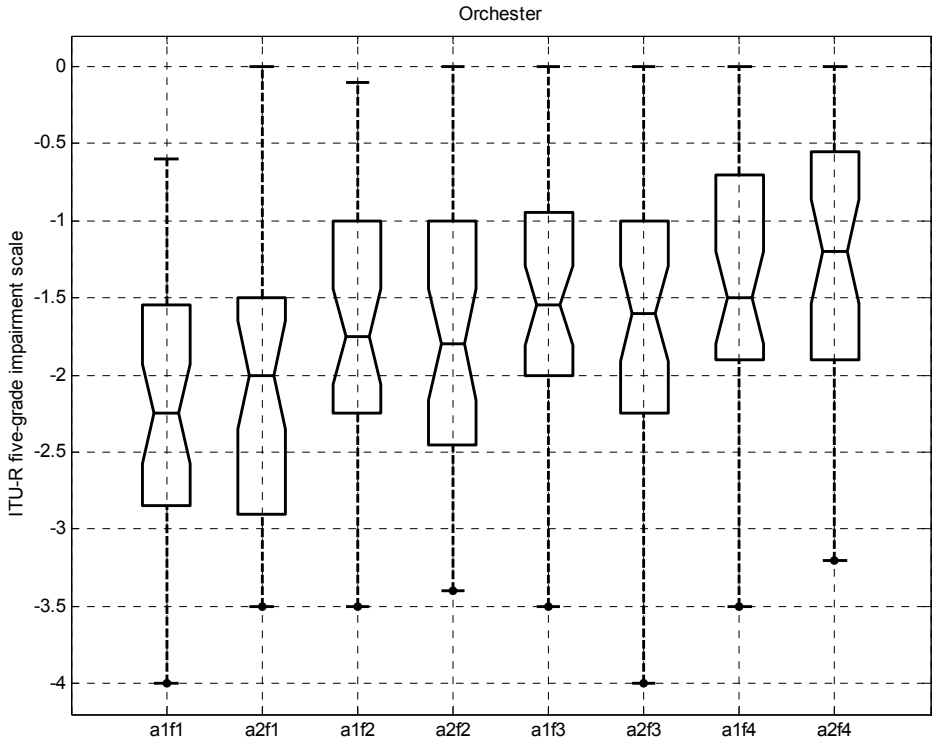

Abb. 4.13: VS1, Audiobeispiel Orchester, Gegenüberstellung der Algorithmen (verschiedene FSs)

Als nächstes betrachtet man das Orchesterbeispiel (vgl. Abb. 4.13). Ansteigend von FS1 bis FS4 lässt sich eine Tendenz der positiveren Bewertung des Signals erkennen. Diese Auswertungen erwecken den Anschein einer Verbesserung mit jedem FS. Zur Erinnerung, die FSs sind normalverteilte Pseudozufallsfolgen (vgl. Abschnitt 4.2.3.1)!
Der Skalenbereich wird von allen VPN gut ausgenutzt. Die Ergebnisse von a1f1 und a1f2 zeigen, dass bei diesen Audiobeispielen keine Fehlbewertungen auftreten. Der größte Unterschied ergibt sich bei FS1 und FS4. Die zwei CIe der Boxplots innerhalb eines FS überlappen sich, der Medianwert der Bewertungen von Alg2 liegt jedoch an der oberen Grenze des CIs von Alg1. Dadurch kann für FS1 und FS4 der Alg1 als der Zweckmäßigere interpretiert werden kann.
Eine generelle Aussage über einen effizienteren Algorithmus zur Verschleierung von DOs in diesem Orchesterbeispiel kann nicht gemacht werden.

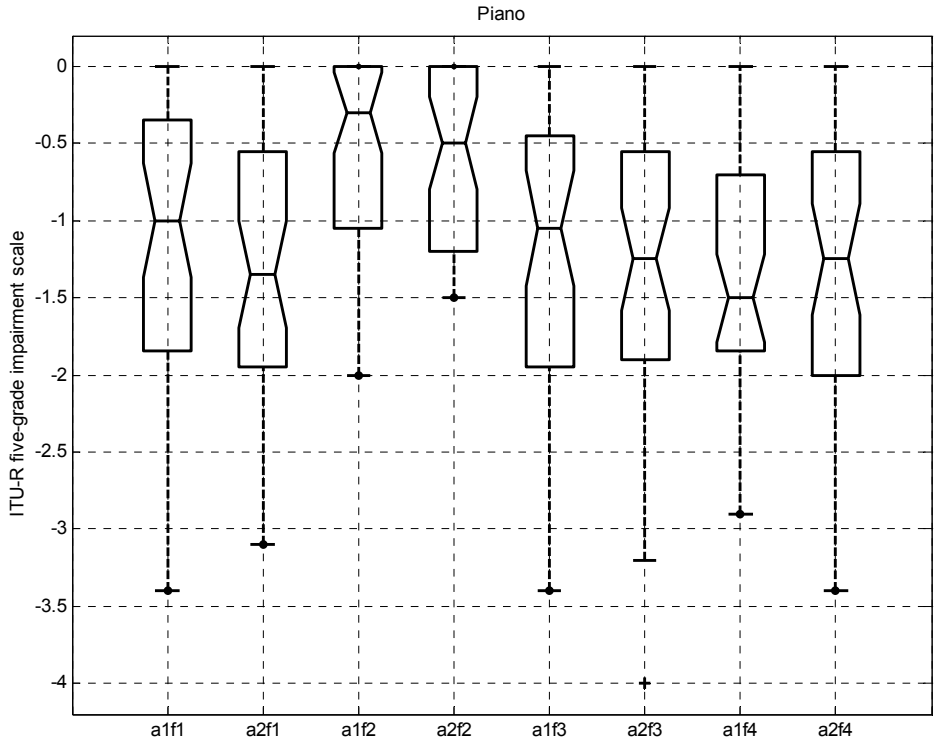

Abb. 4.14: VS1, Audiobeispiel Piano, Gegenüberstellung der Algorithmen (verschiedene FSs)

Beim nächsten Beispiel (Piano, vgl. Abb. 4.14) lassen sich wieder Unterschiede zwischen den Bewertungen der FSs und Algorithmen feststellen. Bis auf FS2 ist bei allen FSs eine gute Ausnutzung der Skala ersichtlich. Da für diese beiden Boxplots $\mathrm{UG}_{FS2}=Q1_{FS2}=0$ gilt, sind 25% ihrer Bewertungen Fehlbewertungen („imperceptible“). Der kleine Skalenbereich $[0,-2]$ lässt auf eine gute Verschleierung der DOs durch die Algorithmen schließen. Weiters liegen 75% aller Bewertungen (von FS1-FS4) im Intervall $[0,-2]$.

Betrachtet man die CIe, lassen sich leichte Differenzen zwischen den Bewertungen der Algorithmen erkennen. Die zwei benachbarten Vertrauensbereiche (innerhalb eines FSs) überschneiden sich für alle FSs. Bei FS1 und FS4 liegen $Q2_{a1f1}$ und $Q2_{a2f4}$ jeweils an der oberen Grenze des benachbarten CIs. Somit ergibt sich eine Tendenz der positiveren Bewertung des Alg1 bei FS1 bzw. des Alg2 bei FS4.

Aus den Ergebnissen geht hervor, dass beide Algorithmen auch bei diesem Beispiel gleichwertig arbeiten. Leichte Unterschiede ergeben sich zwischen den Bewertungen der einzelnen FSs. Über alle FSs betrachtet, kristallisiert sich jedoch kein Algorithmus als der Effizientere heraus. Im Allgemeinen lässt sich eine gute Qualität beider Algorithmen schließen.

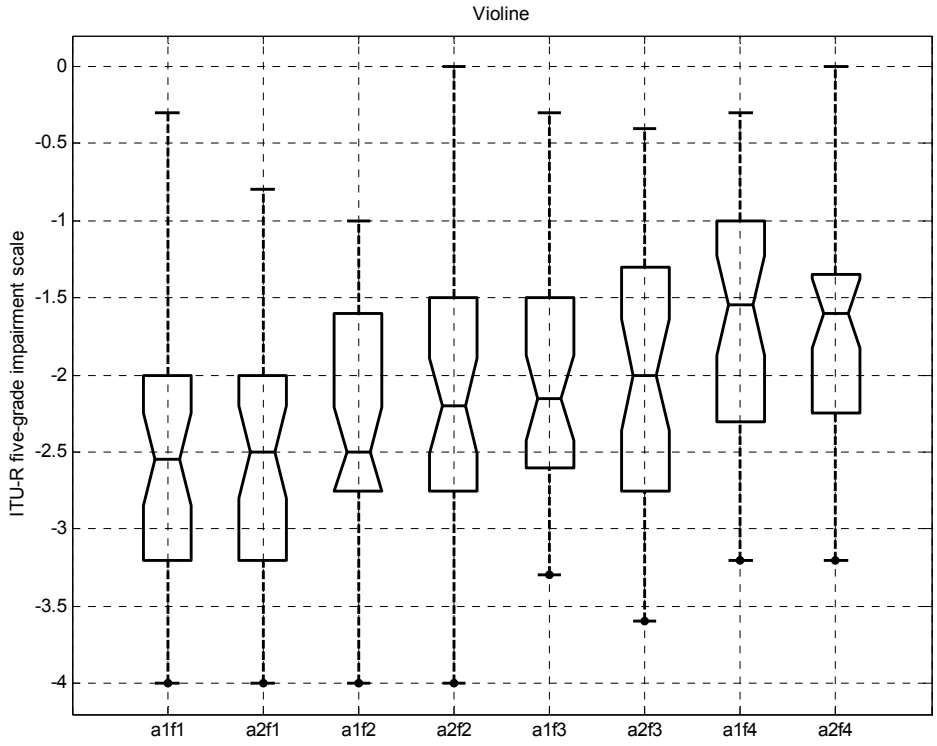

Abb. 4.15: VS1, Audiobeispiel Violine, Gegenüberstellung der Algorithmen (verschiedene FSs)

Die Bewertungen der Violine (vgl. Abb. 4.15) ähneln sehr dem Verlauf der Bewertungen des Orchesters; ansteigend von FS1 bis FS4 wird das Signal immer besser bewertet. Allerdings gibt es deutliche Unterschiede bei den OGn. Beim Orchesterbeispiel treten nur zwei Fälle auf, bei denen $OG \neq 0$ (keine Fehlbewertungen), die Boxplots des Violinebeispiels zeigen hingegen, dass bis auf zwei Fälle (a2f2 und a2f4) $OG \neq 0$ ist. Die VPN haben keine Probleme, die fehlerverschleierten Audiobeispiele von den originalen zu unterscheiden.

Auffallend sind auch die Werte der OGe: für FS1 und FS2 gilt $OG_{FS1} = OG_{FS2} = -4$ („very annoying“), während bei FS3 und FS4 die OGe im Intervall $[-3.2, -3.6]$ liegen. Dies ist eine weitere Bestätigung der Tendenz der besseren Bewertung der FSs.

Nur die Ergebnisse von a2f2 sind über den gesamten Skalenbereich verteilt. Die anderen nutzten einen einseitig (FS1, FS2) bzw. beidseitig (FS3, FS4) beschränkten Bereich. Eine Aussage welcher Algorithmus für welches FS geeigneter ist, kann nur bei FS2 getroffen werden. Aufgrund der Tatsache, dass hier einige Fehlbewertungen auftreten ($UG_{a2f2} = 0$) und $Q2_{a2f2}$ den selben Wert wie der OG des CIs von a1f2 hat, kann Alg2 bei FS2 als der Effizientere angesehen werden; ansonsten sind die Algorithmen gleichwertig.

Die fehlerverschleierten Signale sind bei den Violinebeispielen dem Anschein nach leicht zu erkennen. Daraus folgt eine schlechte Bewertung der Beispiele bzw. der Verschleierungs-algorithmen.

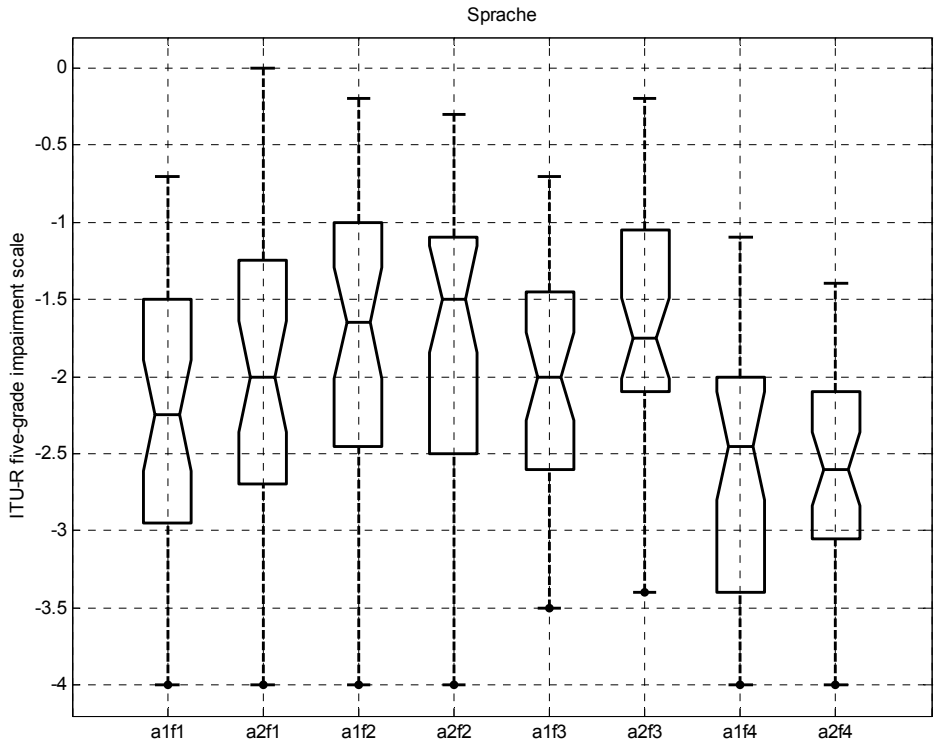

Abb. 4.16: VS1, Audiobeispiel Sprache, Gegenüberstellung der Algorithmen (verschiedene FSs)

Die Bewertung des Sprachbeispiels (vgl. Abb. 4.16) ist nicht wesentlich besser als jene des Violinebeispiels. Nur die Ergebnisse von a2f1 sind über den gesamten Skalenbereich verteilt bzw. nur dort treten Fehlbewertungen auf. Die anderen FSs nutzten einen einseitig (FS1, FS2, FS4; $OG = -4$) bzw. einen beidseitig (FS3) beschränkten Bereich. Daraus folgt eine leichte Detektion der verschleierten DOs in diesem Beispiel.

Durch Betrachtung der Grenzwerte und Quartile der einzelnen Boxplots lässt sich auch einen Tendenz der positiveren Bewertung des Alg1 gegenüber Alg2 feststellen (die CIe überlappen sicht bei jedem FS). Die Ergebnisse des FS4 zeigen, dass dieses FS von allen VPN am schlechtesten (negativsten) bewertet wird.

Sprache ist prinzipiell sehr schwer zu prädizieren, da stochastische Anteile (Rauschen) und nicht stationäre Anteile vorhanden sind. Dies wirkt sich auf die Qualität der Verschleierung aus. Hätte man auch die Sprachverständlichkeit beurteilt, wären die Ergebnisse sicher besser ausgefallen, da die Sprachverständlichkeit bei den Audiobeispielen immer gegeben ist.

4.3.2.2 Fehlbewertungen

Es gibt vier Arten von Fehlbewertungen:

- Bewertung beider Trials mit <0.
- Bewertung von Beispiel B (Originalsignal) anstatt von C (fehlerverschleiertes Signal).
- Bewertung von C (Originalsignal) anstatt von B (fehlerverschleiertes Signal).
- Alle beide werden mit 0 bewertet (Ist bedingt nur in der VS2 gültig!).

Solche Bewertungen werden als „verschleiertes Signal nicht erkannt“ betrachtet und der entsprechende Wert auf 0 („imperceptible“) gesetzt. Diese Selektion ist auch für die VS2 gültig, mit Ausnahme des vierten Falles (vgl. Abschnitt 4.3.3).

Die Gesamtanzahl der Bewertungen pro VD beträgt: $48\ Trials \cdot 20\ VP = 960\ Bewertungen$.

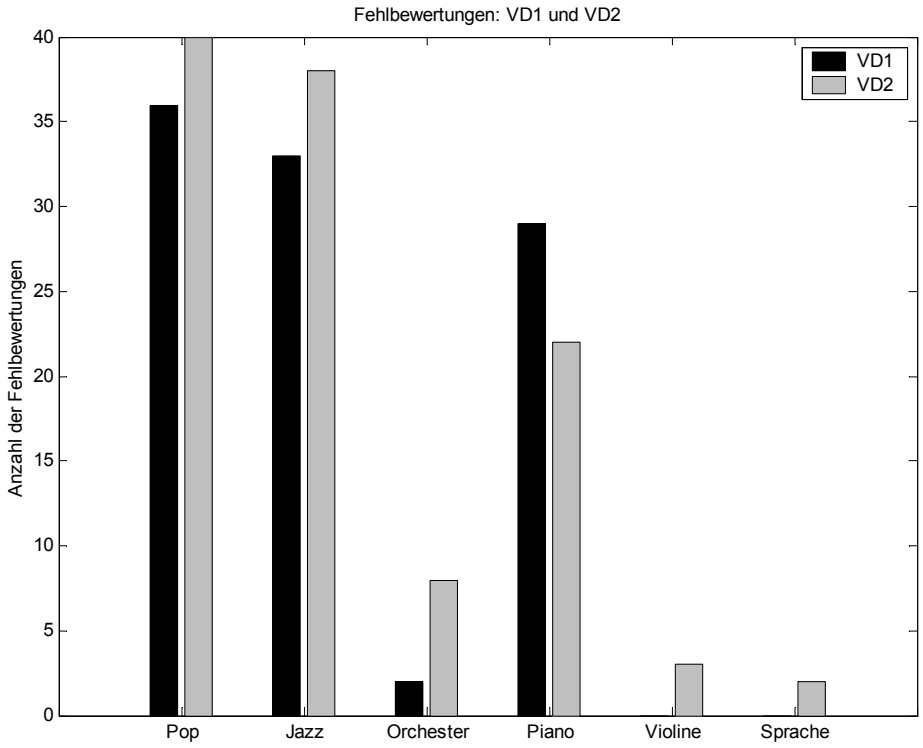

Abb. 4.17: VS1, Fehlbewertungen: VD1 / VD2

Interessant bei Abb. 4.17 bzw. Tabelle 4.5 ist, dass im VD1 das Violine- und das Sprachbeispiel immer richtig detektiert wird, im VD2 jedoch drei- bzw. zweimal nicht. Die höhere Anzahl der Fehlbewertungen im VD2 kann allgemein auf Ermüdungserscheinungen (Konzentrationsmangel) der VPN schließen lassen. Nur beim Pianobeispiel weist der VD1 eine höhere Fehlerquote als VD2 auf. Ein Grund für dieses Ergebnis kann der Lerneffekt der VPN gegenüber der zu bewertenden Artefakte in diesem Beispiel sein: Im VD1 können die

VPN manche fehlerverschleierten Pianobeispiele am Anfang des Versuchs nicht erkennen. Erst durch einige wiedergegebene Beispiele lernen sie, auf gewisse Artefakte zu hören. Der VD1 kann in diesem Fall beinahe als „Einführungstest“ des VD2 gesehen werden.
Aus der Graphik geht auch hervor, dass Artefakte beim Pop-, Jazz- und Pianobeispiel wesentlich schwerer zu detektieren waren, als bei den restlichen drei Beispielen (Orchester, Violine, Sprache). (vgl. Abschnitt 4.3.2.1 Auswertungen und Abb. 4.11 - Abb. 4.16).

Fehlbewertungen				
Audiobeispiel	VD1		VD2	
	[Anzahl]	[%]	[Anzahl]	[%]
Pop	36	16.9	40	18.77
Jazz	33	15.49	38	17.84
Orchester	02	0.94	08	3.76
Piano	29	13.62	22	10.33
Violine	00	0	03	1.41
Sprache	00	0	02	0.94
Summe: 213 (100%)	**100**	**46.95**	**113**	**53.05**

Tabelle 4.5: VS1, Fehlbewertungen

4.3.2.3 Bewertungsstatistik

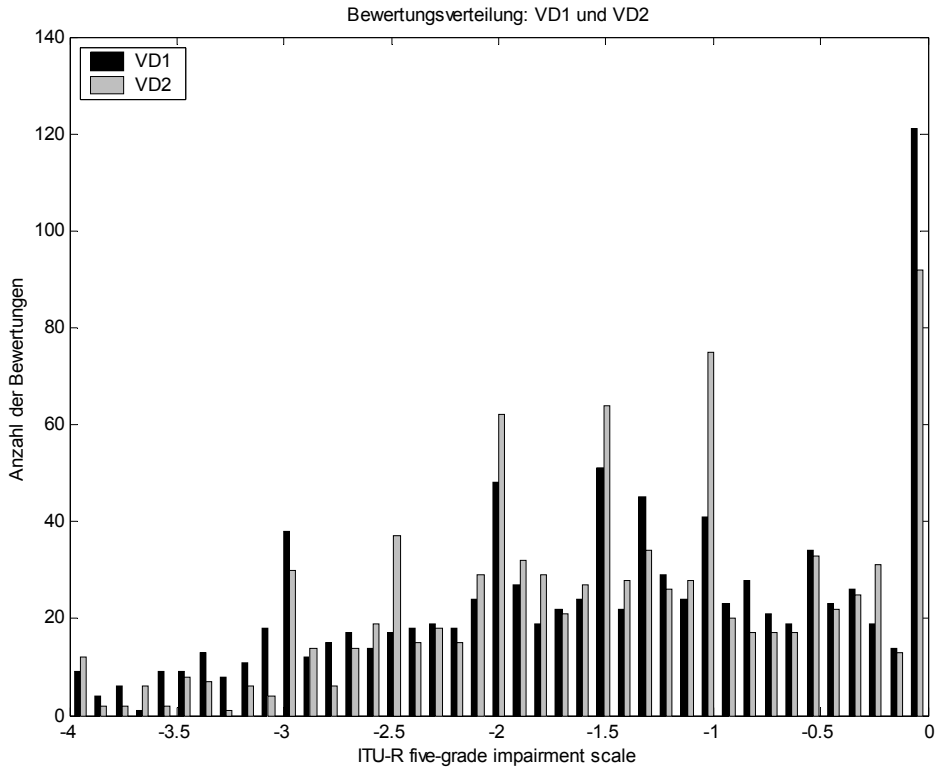

Abb. 4.18: VS1, Verteilung der subjektiven Bewertungen aller VPN

Die Graphik Abb. 4.18 veranschaulicht die Häufigkeitsverteilung der subjektiven Bewertungen aller VPN bezogen auf die ITU-Skala. Im Intervall [0,−2] („imperceptible“-„slightly annoying“) liegen die meisten Bewertungen (72.56%, vgl. Tabelle 4.6).

Intervall	VD1		VD2	
	[Anzahl]	[%]	[Anzahl]	[%]
[0,-2]	680	35.42	713	37.14
]-2,-4]	280	14.58	247	12.86
Summe: 1920 (100%)	**960**	**50**	**960**	**50**

Tabelle 4.6: VS1, Auflistung der Anzahl der Bewertungen

Folglich lässt sich eine Verteilungsfunktion, ähnlich der F-Verteilung, ableiten: gespiegelt um die Ordinate, eine breitgipflige (plateauartige) Verteilung und einem zusätzlichen Höcker am Schluss. Ist das Maß der „skewness“[13] positiv, wird die Verteilung als rechtsschief (oder linkssteil) bezeichnet (dabei gilt: Median > arithmetisches Mittel; ist hier der Fall).
Daraus ergibt sich in Summe eine positive Bewertung der Algorithmen, da die Mehrheit der VPN die Artefakte in den Versuchsbeispielen als „slightly annoying“ bzw. „perceptible, but not annoying“ beurteilt haben. Man erkennt auch die Tendenz der VPN, in ganzen bzw. 0.5er Schritten zu bewerten (vor allem im VD2).
Für einen weiteren (neuen!) subjektiven Hörversuch ist es eine Überlegung wert, die Skala nicht mit Zahlen [0…-4] sonder mit Attributen [unhörbar…sehr störend] zu kennzeichnen. In VS2 bzw. VS3 kann dies noch nicht angewandt werden, da sonst ein Vergleich der Ergebnisse nicht mehr zulässig wäre.

4.3.2.4 Zusammenfassung der VS1

Zusammenfassend lässt sich sagen, dass die beiden Algorithmen, in Abhängigkeit der jeweiligen Audiobeispiele, als gleichwertig anzusehen sind (vgl. Abb. 4.19). Zwischen den Ergebnissen der vier FSs (vgl. Abb. 4.11 bis Abb. 4.16) gibt es einige Fälle, bei denen die Bewertungen des Alg1 oder des Alg2 eine Tendenz der Effizientere zu sein zeigen.
Werden die Ergebnisse der FSs zusammengefasst (vgl. Abb. 4.19), ergibt sich ein ähnliches Resultat. Durch betrachten zweier Boxplots eines Audiobeispiels (z.B. a1b1 und a2b1), erkennt man keine eindeutigen Unterschiede. Die Bewertungsintervalle, in denen die Ergebnisse verteilt sind, bzw. die Grenzwerte und Quartile weisen nur kleine Differenzen auf. Bei den Boxplots der Audiobeispiele fünf und sechs (b5, b6) kann man hingegen eine

[13] Die Parameter der Verteilungsform sind Schiefe („skewness“) und Wölbung („kurtosis“).

Tendenz der besseren Bewertung des Alg2 erkennen: bei den Ergebnissen von Alg1 treten keine Fehldetektioen auf und die CIe sind weiter in den negativen Bereich verschoben.
Aus den genannten Graphiken könnte man auch interpretieren, dass für schwer prädizierbare Signale (Violine, Sprache) keiner der beiden Algorithmen eine brauchbare Audioqualität liefert. Für weitere Ergebnisse (Vergleich zwischen VS1 und VS2) wird auf Abschnitt 4.3.3.4 verwiesen.

Verwendete Abkürzungen in der folgenden Graphik:

axby … „a“ bedeutet Algorithmus, „x“ steht für die entsprechende Nummer (1,2); „b“ steht für Beispiel (Audiobeispiel) und „y“ entspricht wiederum der Nummer (1…6).

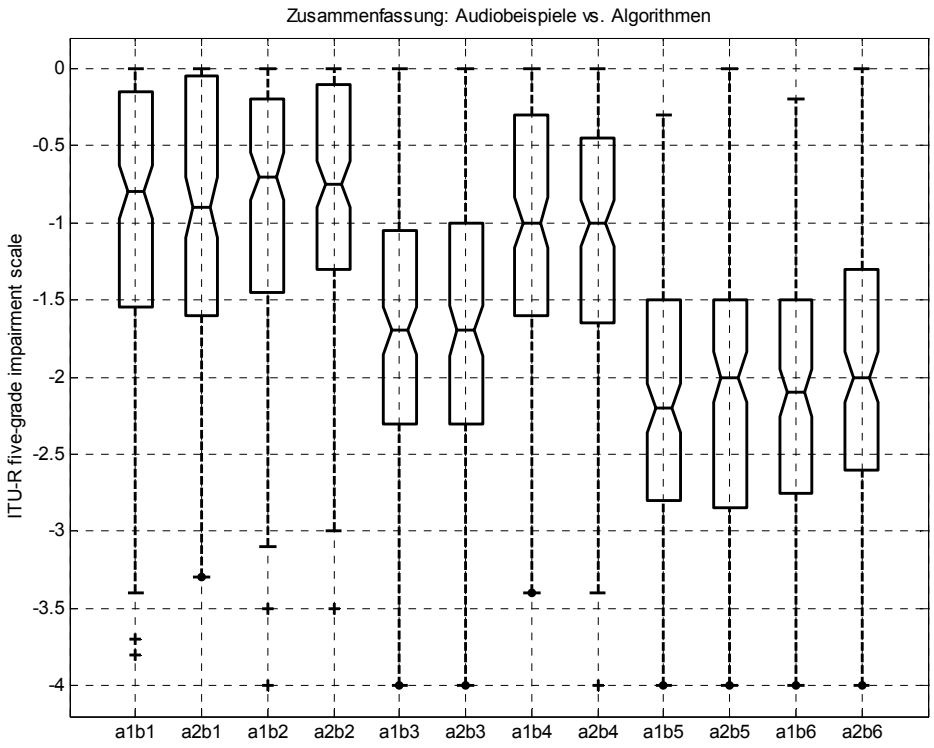

Abb. 4.19: VS1, Audiobeispiele mit Algorithmen

4.3.3 Auswertung der VS2

Die folgenden Diagramme, Abb. 4.20 und Abb. 4.21, stellen die Bewertungen des VD1 bzw. des VD2 der einzelnen Probanden der VS2 dar. Durch Vergleichen der beiden Graphiken erkennt man auch hier eine Tendenz der schlechteren Bewertung des VD2.
Wie bei VS1 gibt es bei VS2 Probanden ohne Fehlbewertungen. Die VP Nr.17 hat sowohl bei VD1 als auch bei VD2 keine Fehlbewertung, die VP Nr.1 nur bei VD2.

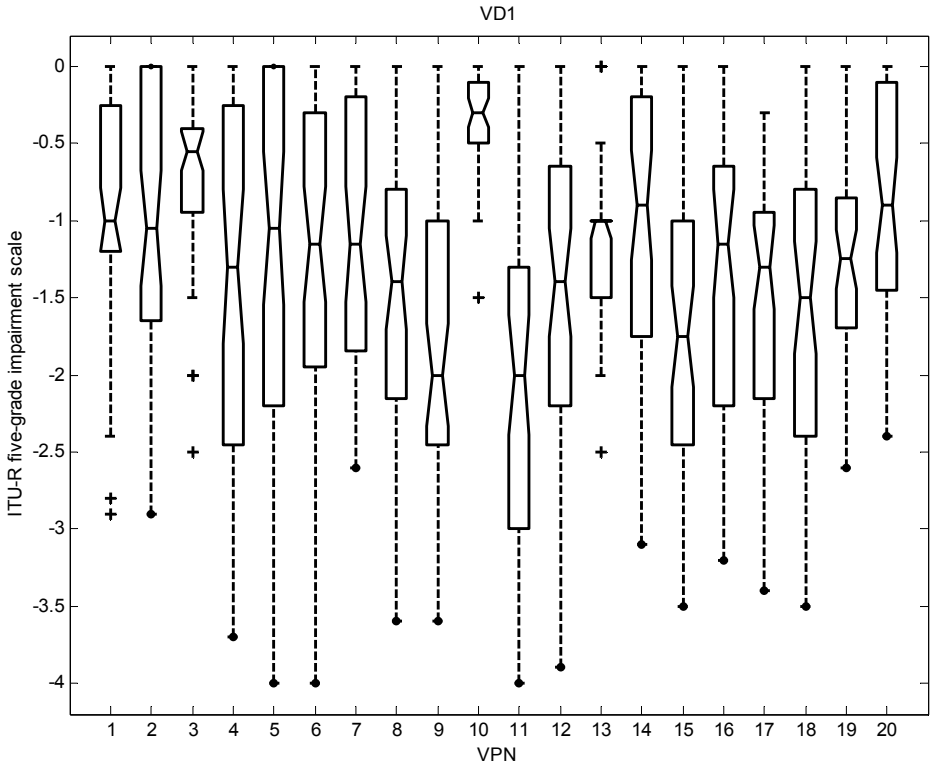

Abb. 4.20: VS2, Bewertung der einzelnen VPN, VD1

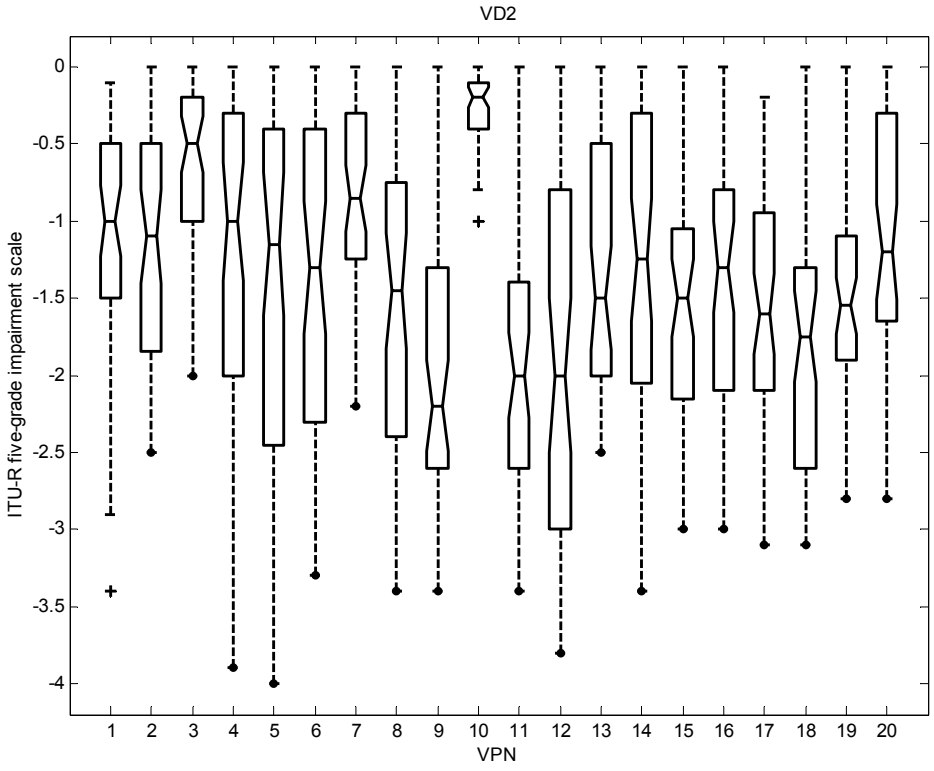

Abb. 4.21: VS2, Bewertung der einzelnen VPN, VD2

Im Vergleich zur VS1 treten nicht so viele Fälle (nur in VD1: VP Nr.2 und Nr.6) mit 25% Fehlbewertungen ($\mathrm{UG} = Q1 = 0$) auf, dafür liegen in VS2 mehr OGe im Intervall $[0, -0.5]$. Die VP Nr.10 hat 75% ihrer Bewertungen in diesem Skalenbereich abgegeben, d.h. drei viertel der Audiobeispiele sind mit „imperceptible“ bewertet worden! (Diese VP ist ein typischer Fall für ein „post-screening“.) Betrachtet man den OG, den berechneten Medianwert und den QA der Boxen der VS1 und vergleicht diese mit denen der VS2, ergibt sich eine positivere Bewertung der VS2.
Bei der Skalenausnützung gibt es keine großen Unterschiede zur VS1. Der Großteil der VPN nutzt die fünfteilige Skala relativ gut aus. Extreme Fokussierungen gibt es bei VP Nr.3, Nr.10 und Nr.13 im VD1 bzw. bei VP Nr.3 und Nr.10 im VD2. Auffällig sind die Bewertungen der VP Nr.13 im VD1. Da $Q1 = Q2 = -1$ ist, hat diese VP 25% der Audiobeispiele mit -1 („perceptible, but not annoying“) bewertet, bzw. 50% der Ergebnisse liegen im Intervall $[-0.5, -1]$.
Trotz einiger Unterschiede zwischen VD1 und VD2 werden die beiden Datensätze wieder für einige Analysen zusammen betrachtet.

4.3.3.1 Auswirkung der FSs

Für die folgenden Graphiken gelten die gleichen Abkürzungen wie in Abschnitt 4.3.2.1. Da dieselben Audiobeispiele und FSs wie in der VS1 für dieVS2 verwendet werden, gelten auch hier die gleichen allgemeinen Schlussfolgerungen, wie sie im vorherigen Abschnitt beschrieben werden.

Die erste Auswertung befasst sich mit der Auswirkung der vier Fehlerszenarios (siehe Abschnitt 4.2.3.1) auf die Bewertung der zwei Algorithmen und der Audiobeispiele. In Abb. 4.22 sind alle Trials eines FS in Abhängigkeit der zwei Algorithmen gegenübergestellt. (Die Graphiken Abb. 7.3 und Abb. 7.4 stellen die Ergebnisse des VD1 und des VD2 getrennt für die Algorithmen und für jedes Audiobeispiel über alle FSs dar.)
Vergleicht man Abb. 4.10 mit Abb. 4.22 zeigt sich, dass sich allgemein die drei Quartilwerte leicht nach oben verschoben haben und damit die Bewertungen insgesamt positiver sind. (Eine Ausnahme ist a2f2; dort stimmen beide Boxplots überein.) Bis auf minimale Unterschiede sind sich alle Boxplots ähnlich.
Aus der Graphik ist auch ersichtlich, dass nicht immer der volle Skalenbereich ausgenutzt wird. In der VS2 werden die fehlerverschleierten Audiobeispiele des FS3 positiver als jene von VS1 bewertet.

Dieses erste Ergebnis hat den Anschein, dass der zweite (veränderte) Einführungstest die VPN effizienter auf den Hörversuch vorbereitet, als der erste.

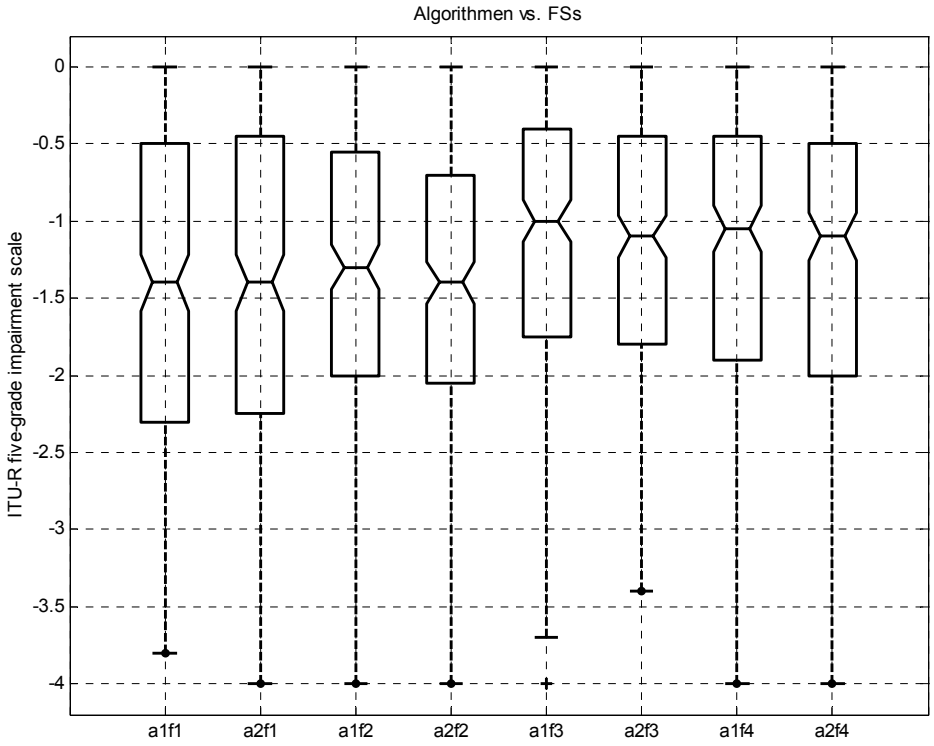

Abb. 4.22: VS2, Gegenüberstellung der zwei Algorithmen mit unterschiedlichen FSs

In den folgenden sechs Graphiken werden wiederum die Bewertungen der einzelnen Audiobeispiele dargestellt (vgl. Abschnitt 4.3.2.1). Entlang der Abszisse sind Alg1 bzw. Alg2 mit den einzelnen FSs (1-4) aufgetragen. Zuerst wird die aktuelle VS (jedes Diagramm) ausgewertet. Anschließend erfolgt der Vergleich der beiden Versuchsreihen miteinander.

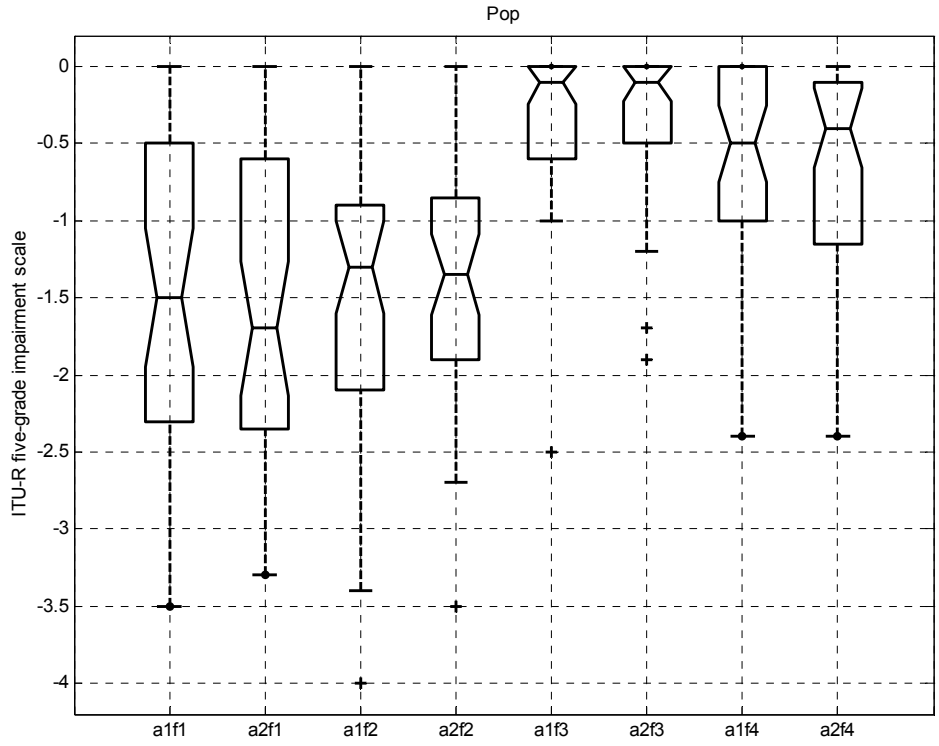

Abb. 4.23: VS2, Audiobeispiel Pop, Gegenüberstellung der Algorithmen (verschiedene FSs)

Die Graphik Abb. 4.23 zeigt die unterschiedlichen Bewertungen des Popbeispiels. Man kann wieder deutlich zwei Bewertungsgruppen unterscheiden. Beide FSs, FS1 und FS2, sind sowohl mit Alg1 als auch mit Alg2 schlechter bewertet worden, als FS3 und FS4.

Ähnlich wie bei VS1, gilt auch hier für FS3 und FS4: $\mathrm{UG} = Q1 = 0$ (mit Ausnahme einer geringen Abweichung von $Q1_{a2f4} = -0.1$), d.h. 25% der Bewertungen sind Fehlbewertungen.

Da die beiden CIe der Boxplots von FS3 mit dem Wert Null beginnen, ist es theoretisch möglich (95%ige Wahrscheinlichkeit), dass $Q2_{a1f3} = Q2_{a2f3} = 0$; 50% der Bewertungen wären Fehlbewertungen. Die Ergebnisse von FS3 weisen gegenüber jenen der VS1 eine noch stärkere Einschränkung des verwendeten Skalenbereichs auf, da 75% der Bewertungen im Intervall $[0, -0.6]$ bzw. 100% der Bewertungen im Intervall $[0, -1.2]$ („imperceptible“, „perceptible, but not annoying“) liegen.

Abgesehen von FS1 ist der Trend der positiveren Bewertung der VS2 (vgl. Abb. 4.22 mit Abb. 4.10) auch hier erkennbar; am deutlichsten bei FS3. Wie in VS1 gilt für die Ergebnisse der VS2, dass das Popbeispiel mit FS3 und FS4 wesentlich besser bewertet wird als mit FS1 und FS2.

Durch Vergleichen aller Ergebnisse kann kein Algorithmus für die Verschleierung von DOs in einem Popbeispiel favorisiert werden; beide Algorithmen arbeiten gleichwertig.

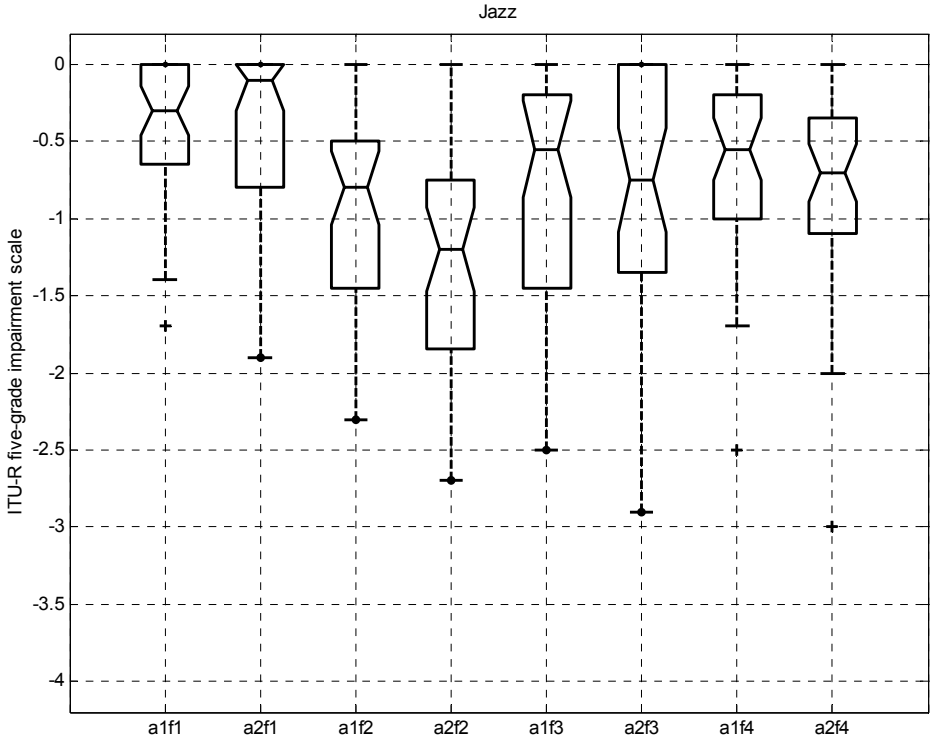

Abb. 4.24: VS2, Audiobeispiel Jazz, Gegenüberstellung der Algorithmen (verschiedene FSs)

Als nächstes wird das Jazzbeispiel (Abb. 4.24) betrachtet. Es gibt drei Fälle, bei denen 25% der Bewertungen Fehlbewertungen sind: a1f1, a2f1 und a2f3 ($UG = Q1 = 0$), bzw. einen Fall, bei dem 50% Fehlbewertungen vorkommen können (a2f1).

Alle Ergebnisse der FSs sind analog zur VS1 in einem sehr eingeschränkten Skalenbereich verteilt. Bis auf jenen von FS3 sind die Skalenbereiche durch die Ergebnisse der VS2 noch weiter eingeschränkt. Die verminderte Anzahl der Ausreißer lässt auf einen „einheitlichen“ Bewertungsbereich der VPN schließen; man könnte auch sagen, die VPN sind sich untereinander einig über ihre Bewertungen.

Weitere Unterschiede zwischen den Bewertungen der beiden VSN ergeben sich durch Vergleichen der berechneten Medianwerte bzw. der CIe der VS1 (vgl. Abb. 4.12) mit jenen von VS2. Bei den Boxplots von VS2 gibt es wesentlich größere Differenzen zwischen den Werten von *Q2* bzw. vom CI innerhalb eines FSs, wogegen diese Werte bei VS1 innerhalb eines FSs nur geringfügig voneinander abweichen (größere Abweichung bei FS2).

Bei VS1 können die beiden Algorithmen als gleichwertige betrachtet werden; tendenziell wird bei VS2 die Anwendung von Alg1 besser eingestuft als jene von Alg2. Ein Vergleich zur VS1 liefert insgesamt wiederum eine positivere Bewertung des Beispiels.

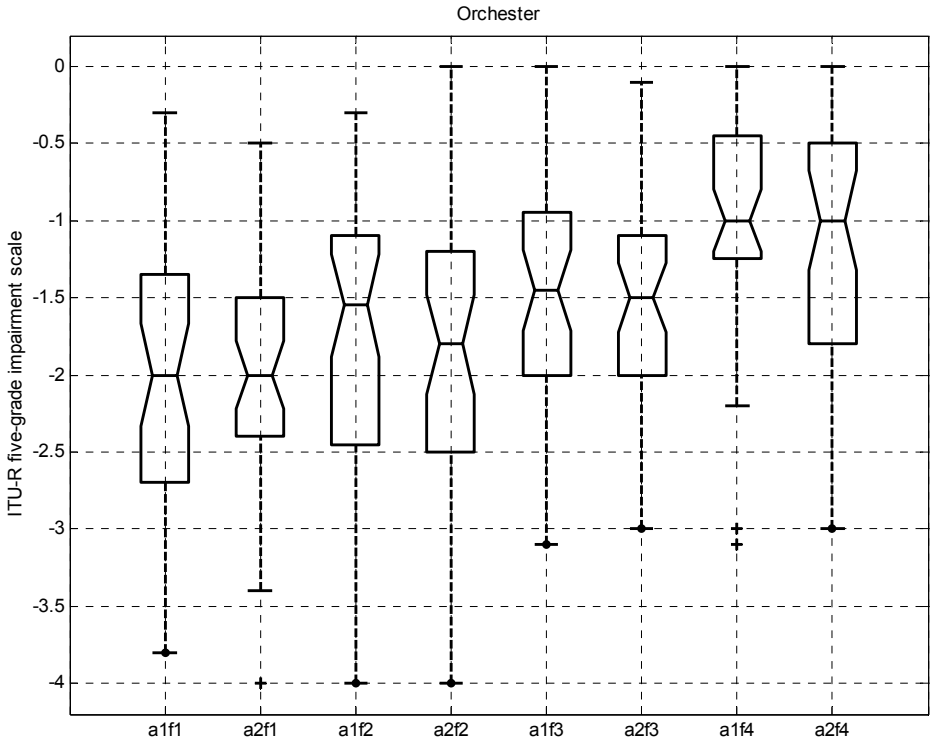

Abb. 4.25: VS2, Audiobeispiel Orchester, Gegenüberstellung der Algorithmen (verschiedene FSs)

Das Diagramm in Abb. 4.25 zeigt die Auswertungen für das Orchesterbeispiel. Man erkennt auch hier die Tendenz einer positiveren Bewertung, ansteigend von FS1 bis FS4. Diesmal ändern sich jedoch die beiden Grenzwerte und somit die Bewertungsbereiche stärker als in Abb. 4.13. Es gibt vier Fälle, bei denen die fehlerverschleierten Audiobeispiele leicht zu detektieren sind, da keine Fehlbewertungen auftreten (a1f1, a2f1, a1f2 und a2f3). Die UGe des FS3 und FS4 sind deutlich positiver, jene von FS2 deutlich negativer als die äquivalenten in VS1.

Im Vergleich zu VS1 überlagern sich die CIe von FS1 und FS4 nun mehr, die Medianwerte stimmen gänzlich überein. Das bedeutet, dass in diesen zwei Fällen praktisch kein Unterschied zwischen den Algorithmen feststellbar ist. Nur bei der Auswertung von FS2 zeigt einen geringfügigen Vorteil von Alg1; ansonsten arbeiten die Algorithmen qualitativ gleich.

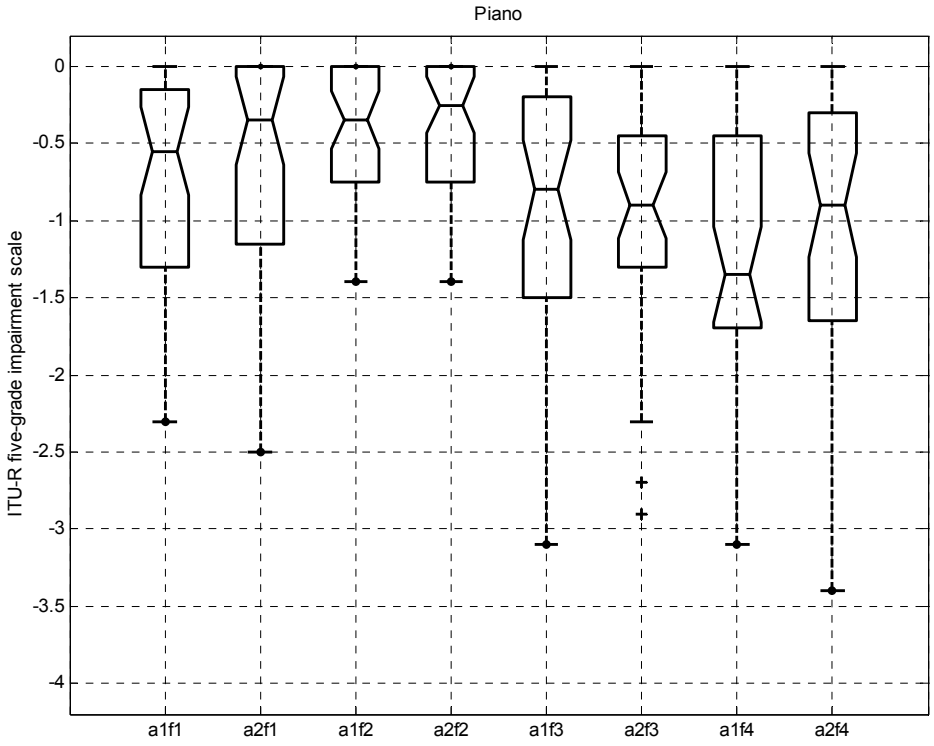

Abb. 4.26: VS2, Audiobeispiel Piano, Gegenüberstellung der Algorithmen (verschiedene FSs)

Die Bewertungen des Pianobeispiels der VS2 sind in Abb. 4.26 dargestellt. Bei den Ergebnissen von drei fehlerverschleierten Audiobeispielen treten 25% Fehlbewertungen auf (a2f1, a1f2 und a2f2). Über alle Boxplots gesehen, haben sich im Vergleich zur VS1, die UGe Richtung Null verschoben. Weiters erkennt man eine kleinere Ausnutzung des Skalenbereichs durch die Ergebnisse der VS2. Die Boxplots des FS2 zeigen ganz deutlich, dass sie die Skala noch weniger als in VS1 ausnutzten.

Wie beim Orchesterbeispiel sind die Unterschiede zwischen den Algorithmen geringer geworden; die Bewertungen sind etwas homogener. Der einzige signifikante Unterschied ergibt sich bei FS4. Dort wird Alg1 besser bewertet (die CIe von a1f4 und a2f4 überschneiden sich nur sehr wenig).

Zusammenfassend kann man bei diesem Audiobeispiel auf eine weitaus positivere Bewertung der VS2 schließen. Über FS1, FS2 und FS3 lassen sich keine konkreten Aussagen über die effizientere Anwendung eines Algorithmus machen, bei FS4 kann Alg2 favorisiert werden.

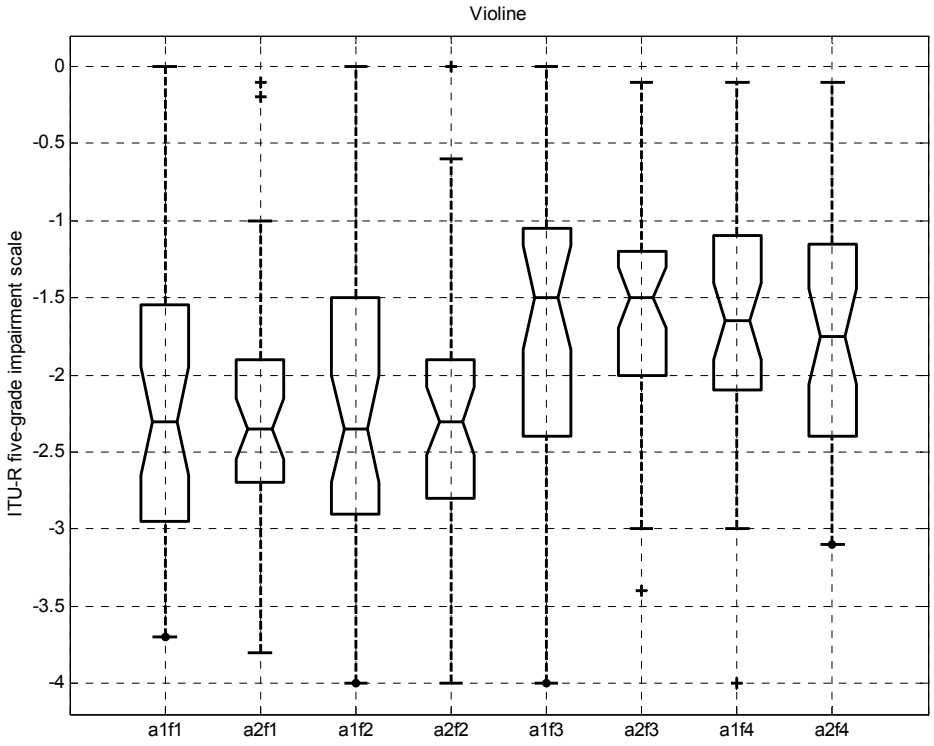

Abb. 4.27: VS2, Audiobeispiel Violine, Gegenüberstellung der Algorithmen (verschiedene FSs)

Man kann in Abb. 4.27 (Violine) wieder zwei Gruppen erkennen (FS1 und FS2 bzw. FS3 und FS4); ähnlich der Bewertung des Popbeispiels in VS2 (vgl. Abb. 4.23). Allerdings sind die Boxplots nach unten (in den negativeren Bereich) verschoben. Bei den Ergebnissen der VS1 lässt sich hingegen eine Tendenz der positiveren Bewertung, ansteigend von FS1 bis FS4 feststellen.

Auffallend sind die fast vollkommen identen Medianwerte bzw. CIe der Ergebnisse der VS2 innerhalb eines FSs. Weiters treten bei a2f1, a2f2 (bis auf einen Ausreißer), a2f3, a1f4 und a2f4 keine Fehlbewertungen auf. Für die ersten drei Fälle sind die Werte von QA kleiner und die OGe positiver (ausgenommen OG_{a1f3}) als jene der VS1. Aus diesen Tatsachen lässt sich schließen, dass für diese Beispiele Alg1 der effizientere ist.

Im Vergleich zur VS1 erkennt man auch positivere UGe (mit Ausnahme der Werte von a1f2 und a2f2). Die VPN der VS2 können die fehlerverschleierten Audiosignale nicht mehr so leicht detektieren oder empfinden die Artefakte als nicht so störend, wie die Probanden der VS1.

Für das Violinbeispiel lässt sich, trotz seiner eher schlechten Bewertung, eine Tendenz der positiveren Bewertung der VS2 ablesen. Bei FS1, FS2 und FS3 ist die Bewertung des Alg1 positiver.

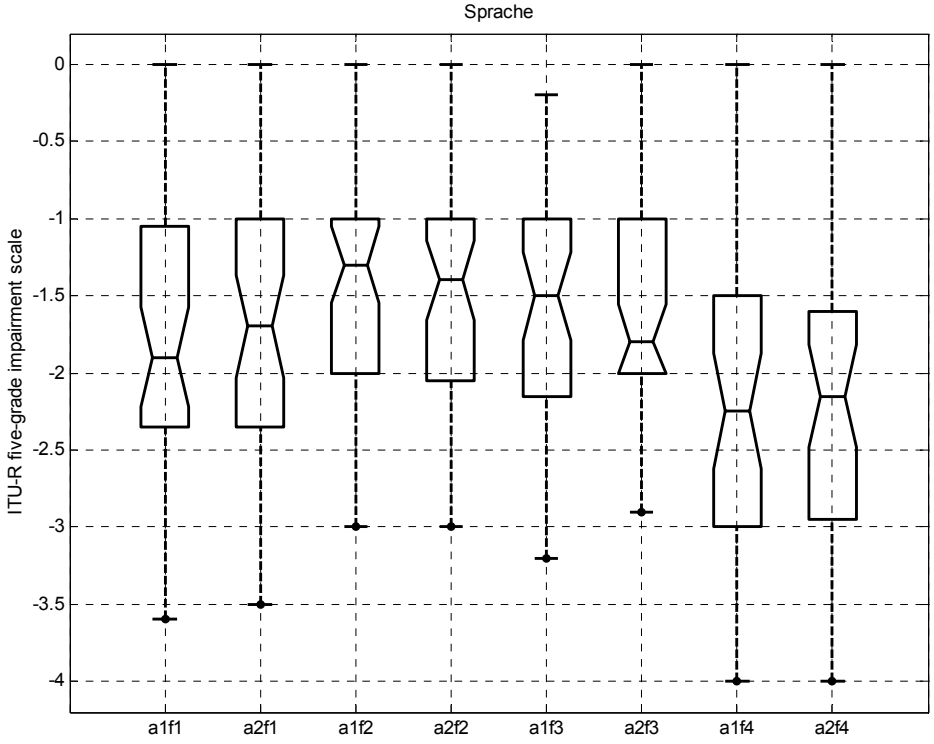

Abb. 4.28: VS2, Audiobeispiel Sprache, Gegenüberstellung der Algorithmen (verschiedene FSs)

Das Diagramm der Auswertungen des Sprachbeispiels (vgl. Abb. 4.28) unterscheidet sich vollkommen von jenem der VS1 (vgl. Abb. 4.16). Da für alle Boxplots der VS2 UG=0 gilt (bis auf $UG_{a1f3} = -0.2$), deutet dies auf vermehrte Fehlbewertungen (25%) hin. Bei VS1 gibt es nur einen einzigen Fall (a2f1) mit Fehlbewertungen.

Vergleicht man die OGe der beiden VSN, bekommt man folgendes Ergebnis: bei VS1 haben sechs von acht OGe den Wert -4, bei VS2 jedoch nur zwei von acht. Auffallend ist auch der konstante Wert $Q1 = -1$ für die ersten drei FSs. Der gesamte Bewertungsbereich hat sich bei VS2 nach oben, Richtung Null, verschoben. Dies bestätigt wieder die positivere Bewertung der VS2 gegenüber VS1.

Bei FS3 ist der signifikanteste Unterschied zwischen den VSN und den Bewertungen der Algorithmen feststellbar. Aus den Ergebnissen der VS1 ergibt sich für a2f3 ein $Q2$, der mit dem oberen Wert des CIs von a1f3 zusammenfällt, d.h. Alg2 wird hier positiver bewertet. Aus den Boxplots der VS2 kann mit der gleichen Überlegung Alg1 als der bessere eruiert werden. Am deutlich schlechtesten werden in beiden VSN die Audiobeispiele des FS4 bewertet.

Aus den Auswertungen können keine eindeutigen Aussagen über effizienteren Algorithmus getroffen werden.

4.3.3.2 Fehlbewertungen

Eine Auflistung, wann eine Bewertung als Fehlbewertung gewertet wird, findet man unter Abschnitt 4.3.2.2.

Die Gesamtanzahl der Bewertungen pro VD beträgt: $52\ Trials \cdot 20\ VP = \underline{\underline{1040\ Bewertungen}}$;

ohne Placebobeispiele ergeben sich wieder 960 Bewertungen.

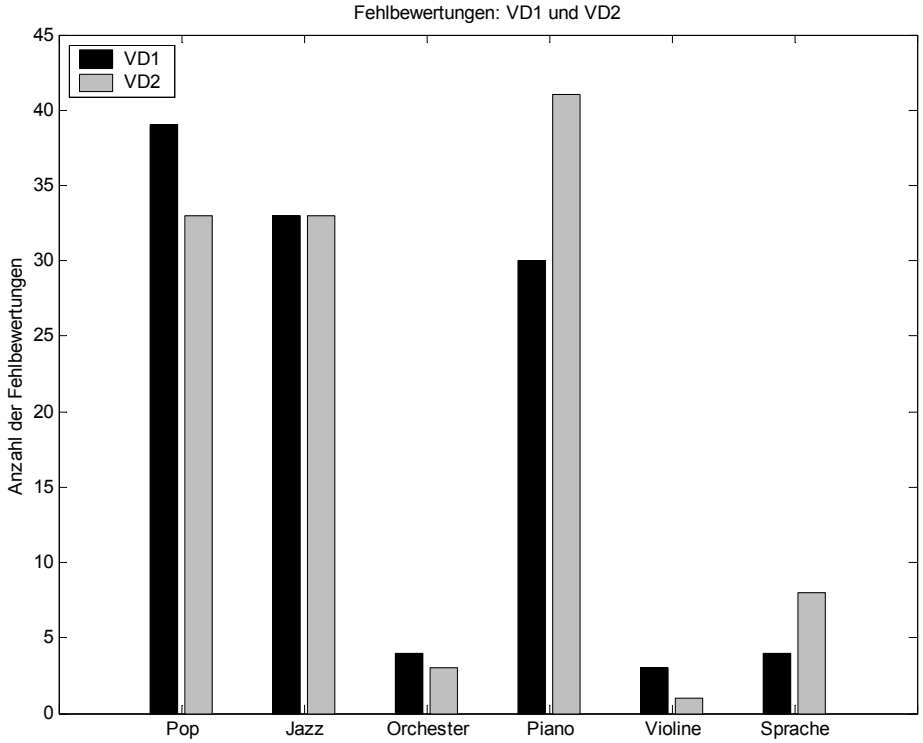

Abb. 4.29: VS2, Fehlbewertungen: VD1 / VD2

Fehlbewertungen				
Audiobeispiel	VD1		VD2	
	[Anzahl]	[%]	[Anzahl]	[%]
Pop	39	16.81	33	14.22
Jazz	33	14.22	33	14.22
Orchester	04	1.73	03	1.29
Piano	30	12.93	41	17.68
Violine	03	1.29	01	0.43
Sprache	04	1.73	08	3.45
Summe: 232 (100%)	**113**	**48.71**	**119**	**51.29**

Tabelle 4.7: VS2, Fehlbewertungen

Die Fehlbewertungen des VD1 (vgl. Tabelle 4.7) entsprechen in etwa jenen der VS1. Der VD2 weist gröbere Unterschiede auf. Bei den ersten drei Audiobeispielen treten nicht so oft Fehlbewertungen auf. Mit 17.68% Fehlbewertungen wird das Pianobeispiel im Vergleich zu

vorher fast doppelt so oft falsch bewertet. Auch bei der Sprache ergibt sich eine dreimal höhere Fehlerquote als bei der VS1.

Der Grund dafür könnte auf einen gesteigerten Konzentrationsmangel (Ermüdungserscheinungen) der VPN aufgrund der längeren Versuchsdauer (der Einführungstest dauert doppelt so lange, vier Placebobeispiele) zurückzuführen zu sein.

Beim Pop-, Orchester- und Violinebeispiel treten im VD1 mehr Fehlbewertungen auf als im VD2. Eine mögliche Erklärung dieses Ergebnisses ist, dass die VPN die Artefakte in diesen Beispielen trotz des Einführungstests nicht bzw. sehr schwer wahrgenommen haben. Erst im VD2 können sie kleine Unterschiede detektieren. Der VD1 hat somit die Funktion des Einführungstests für VD2 übernommen. Das Jazzbeispiel weist eine gleiche Fehlerquote bei VD1 und VD2 auf.

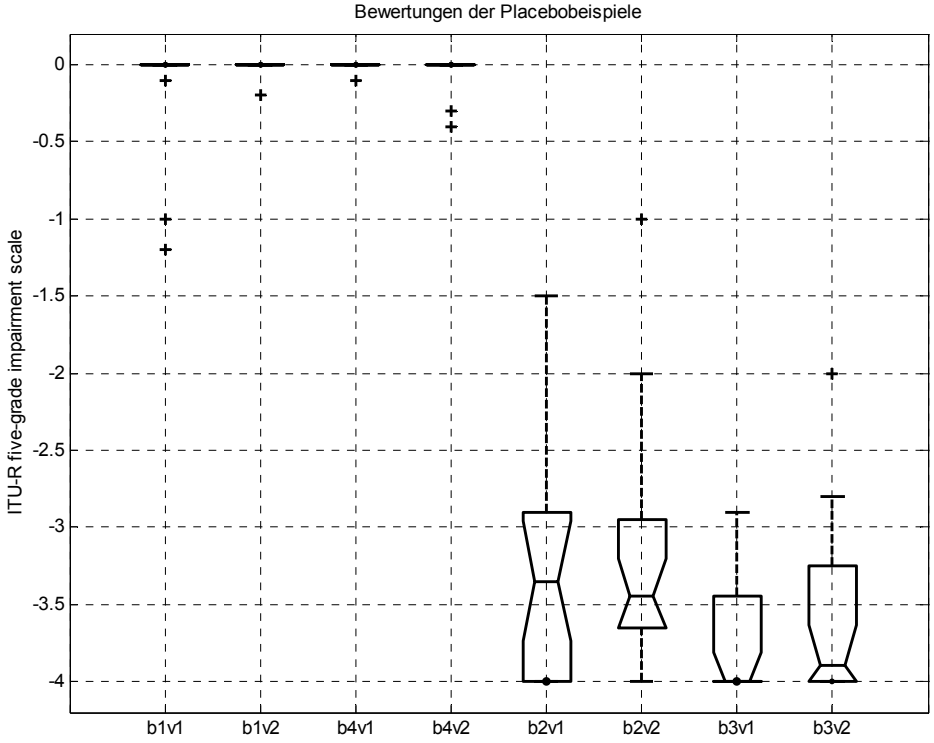

Abb. 4.30: VS2, Bewertung der Placebobeispiele (VD1 / VD2)

Das Diagramm Abb. 4.30 zeigt die Gegenüberstellung der Bewertungen der Placebobeispiele des VD1 bzw. des VD2. (Zur Erinnerung: Den VPN ist bekannt, dass Placebobeispiele vorkommen und dass sie einerseits Originalsignale bzw. andererseits unverschleierte, mit Ausfällen behaftete Signale sein können!)

Die Originalbeispiele werden von allen Probanden (bis auf 7 von 20) ohne Probleme erkannt und dementsprechend bewertet. Bei den Beispielen mit den unverschleierten DOs gibt es

interessanterweise große Abweichungen. Die CIe des Jazzbeispiels liegen im Intervall $[-2.95, -3.75]$, während beim Orchesterbeispiel 50% der Bewertungen mit -4 („very annoying“) erfolgen. Mit Hilfe der UGe wird ersichtlich, dass beim Jazzbeispiel Signalausfälle weniger kritisch beurteilt werden als beim Orchesterbeispiel. Diese Werte und die zwei Ausreißer (-1,-2) zeigen, dass für manche VPN Audiobeispiele mit unverschleierten DOs qualitativ jenen mit schlecht verschleierten DOs entsprechen.

Vergleicht man die Bewertungen des Jazz- und Orchesterbeispiel mit unverschleierten DOs mit den Bewertungen der VS1 bzw. VS2 (vgl. Abb. 4.12, Abb. 4.13, Abb. 4.24, Abb. 4.25), lässt sich ein Unterschied zwischen schlecht verschleierten Signalen und Signalen mit DOs feststellen, d.h. die Signale mit den unverschleierten DOs werden definitiv schlechter bewertet.

Ruft man sich die niedrigen Audioqualitätsansprüche vieler Personen ins Gedächtnis und betrachtet dann die Boxplots der Placebobeispiele mit den unverschleierten DOs, so überrascht dieses Ergebnis nicht allzu sehr.

Bei den ersten vier Boxplots fallen die drei Quartilwerte, der UG und der OG zusammen (aufgrund der schwarz weiß Graphik kann man sie nicht unterscheiden).

4.3.3.3 Bewertungsstatistik

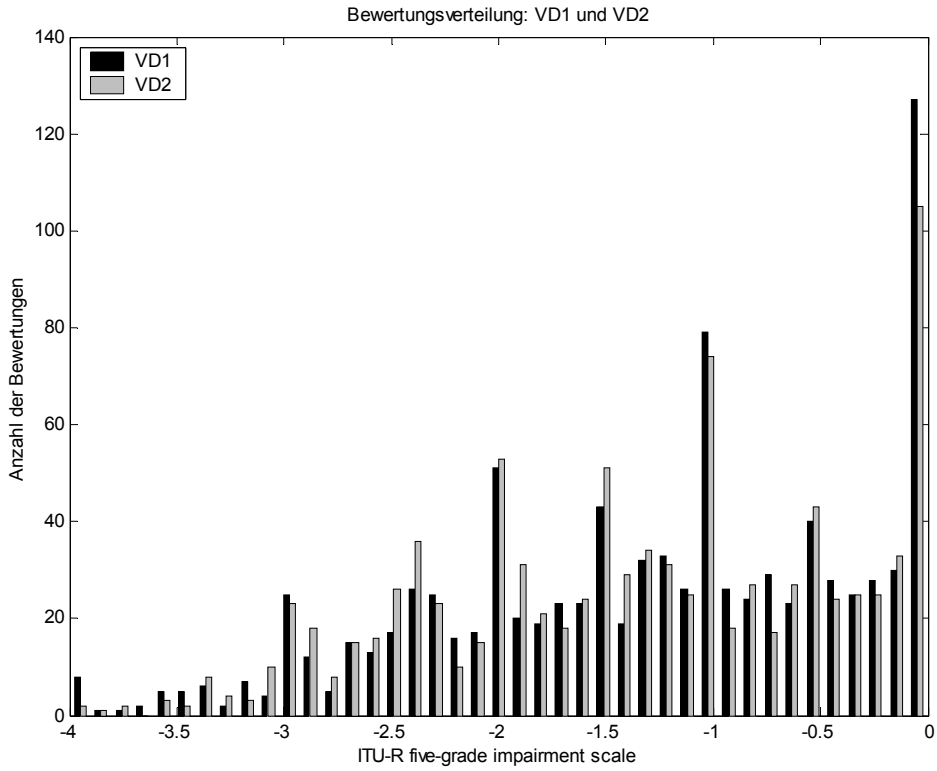

Abb. 4.31: VS2, Verteilung der subjektiven Bewertungen aller VPN

Abb. 4.31 veranschaulicht die Anzahl der eingegebenen Bewertungen aller VPN entsprechend der ITU-Skala. Mit Hilfe der Graphik und der Tabelle 4.8 wird ersichtlich, dass im Vergleich zur VS1 (vgl. Abb. 4.18, Tabelle 4.6) nicht wesentlich mehr Bewertungen (77.24%) im Intervall [0,-2] liegen. Man kann auch bei dieser Bewertungsverteilung auf eine rechtssteile Verteilung schließen (vgl. Abschnitt 4.3.2.3).

Intervall	VD1		VD2	
	[Anzahl]	[%]	[Anzahl]	[%]
[0,-2]	748	38.96	735	38.28
]-2,-4]	212	11.04	225	11.72
Summe: 1920 (100%)	**960**	**50**	**960**	**50**

Tabelle 4.8: VS2, Auflistung der Anzahl der Bewertungen

In Summe ergibt sich eine positive Bewertung der Algorithmen, da die Mehrheit der VPN die Artefakte in den Versuchsbeispielen als „slightly annoying“ bzw. „perceptible, but not annoying“ beurteilt hat. Die Tendenz der VPN, in ganzen bzw. 0.5er Schritten zu bewerten ist bei beiden Durchläufen wesentlich stärker ausgeprägt als bei der VS1.

4.3.3.4 Zusammenfassung der VS2 (und VS1 vs. VS2)

Zusammenfassend lässt sich das selbe wie unter Abschnitt 4.3.2.4 sagen. Die beiden Algorithmen, über die Bewertungen aller Audiobeispiele und FSs betrachtet, sind als gleichwertig anzusehen (vgl. Abb. 4.32). Es treten nur minimale Bewertungsunterschiede (Tendenzen) in Bezug auf FSs bzw. Audiobeispiele auf. Für schwer prädizierbare Signale (Violine, Sprache) liefert keiner der beiden Algorithmen eine brauchbare Audioqualität.

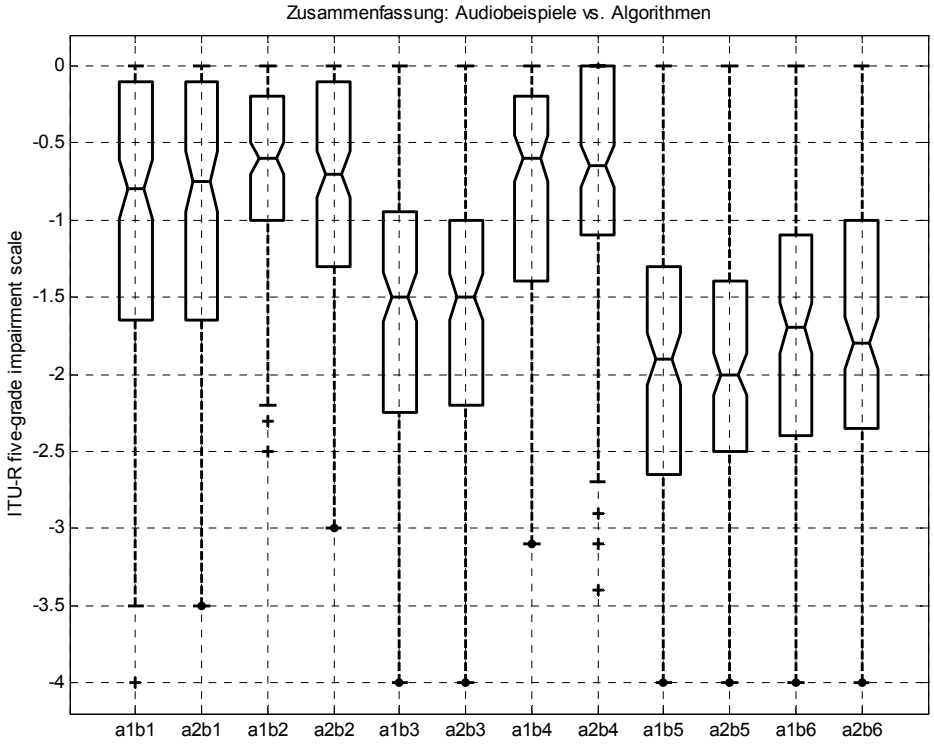

Abb. 4.32: VS2, Gegenüberstellung: Audiobeispiele / Algorithmen (über alle FSs)

Durch Vergleicht der Abb. 4.19 mit Abb. 4.32 erkennt man, dass bei VS2 bei jedem Audiobeispiel Fehlbewertungen auftreten (bei VS1 gibt es zwei Beispiele ohne Fehlbewertungen: a1b5 und a1b6). Ein signifikanter Unterschied ergibt sich bei b4. Die CIe sind um 0.4 positiver geworden bzw. $QI_{a2b4} = 0$, d.h. es sind 25% der Bewertungen Fehlbewertungen.

Die mittels Alg1 berechneten fehlerverschleierten Violine- und Sprachbeispiele weisen eine leichte positivere Bewertung gegenüber VS1 auf. Weiters hat sich der OG_{a1b2} verbessert.

Der Vergleich der beiden VSN zeigt die bessere Bewertung von VS2 gegenüber VS1. Der ausschlaggebende Grund dafür war sicher die verlängerte (intensivere) Einführungsphase jeder VP. Im Nachhinein gesehen wäre es sinnvoller gewesen, den Einführungstest noch ausführlicher zu gestalten und dafür nur einen VD durchzuführen. Dadurch kann auf die zufällige Auswahl der Trials verzichtet werden, da die Gefahr des Lerneffekts[14] für den VD2 nicht mehr gegeben ist.

Da beide Algorithmen im Gesamten als gleichwertig angesehen werden können, ergeben sich automatisch Untersuchungen der Recheneffizienz der Algorithmen. Diese werden jedoch nicht in dieser Arbeit behandelt.

[14] Die VP merkt sich die Reihenfolge ihrer Bewertungen und bewertet den VD2 gleich wie den VD1.

5 Dritte Versuchsserie (VS3)

Dieses Kapitel befasst sich mit der Auswertung der VS3. Die beiden beschriebenen Algorithmen, ZCR und AMDF, werden in den Interpolationsalgorithmus (vgl. Abschnitt 2.1.1.2) implementiert, mit ihnen diverse Testbeispiele erstellt und anschließend mittels Hörversuch auf ihre Effizienz getestet.

5.1 Allgemeines zur VS3

Die VS3 wird im Vergleich zu VS2 nicht wesentlich verändert. Von den insgesamt 21 Probanden sind elf (52.38%) als „expert listeners“ anzusehen. Der relativ hohe Prozentsatz an qualifizierten Hörern lässt auf ein kritischeres, aussagekräftiges Ergebnis schließen.
Da bis auf vier Testpersonen (19.05%) jede VP die VS1 oder die VS2 schon durchgeführt hat, kann auf den Einführungstest verzichtet werden. Jene vier Probanden erhalten eine längere und detailliertere Erklärung des Versuchsablaufs.

Toningenieur-Studierende	09 (1)
Personen mit „musikalischer Erfahrung“	02 (1)
Personen ohne „musikalische Erfahrung“	10 (4)
Probanden	**21 (6)**

Die Zahlen in Klammer sagen aus, wie viele davon Frauen sind. Der Anteil an weiblichen VPN beträgt 28.57%. Von den einundzwanzig Probanden sind vier Probanden über dreißig Jahre alt, das entspricht 19.05%.
Durchgeführt wird der subjektive Bewertungstest wieder am IEM im Experimentalstudio. Als Wiedergabeart wird die Kopfhörerwiedergabe verwendet.

Modifikationen (Audiobeispiele, Algorithmen und FSs):

Beim Entwerfen der VS3 ist ein Hauptkriterium die Dauer des Hörversuchs gewesen. Je kürzer der Versuch dauert, desto weniger treten Ermüdungserscheinungen auf. Weiters hat man kein Problem mit dem Anwerben von Probanden, da sich die Personen eher 10-20min Zeit nehmen, als über eine Stunde (vgl. VS1 und VS2!).

Mit Hilfe der Ergebnisse der VS1 und VS2 ist es möglich, die ursprüngliche Anzahl von sechs Audiobeispielen auf drei zu reduzieren. Dabei wird auf eine Gleichverteilung der Qualitäten der Beispiele geachtet. Schlussendlich ergeben sich drei Audiobeispiele, die bei den vorherigen Hörversuchen von sehr gut bis eher mittelmäßig bewertet werden. Die Bezeichnung der Testbeispiele stimmt mit den ursprünglichen nicht mehr überein:

AB1.wav … Blackbird – „Grey – The World” – by Heinrich von Kalnein

AB2.wav … Ludwig van Beethoven, Symphonie Nr.5 c-moll op. 67, Ouvertüre „Leonore II”, Wiener Philharmoniker – „Allegro”

AB3.wav … EBU: tec_sqam_08m_bwf_tcm6-12488 (Violine)

Für jede VP kann die gleiche Reihenfolge der wiedergegebenen Trials verwendet werden, da nur ein VD durchgeführt wird und somit die Gefahr des Lerneffekts für den VD2 nicht mehr gegeben ist.

Reihenfolge der Trials: [22 16 12 9 5 13 4 17 2 10 6 20 11 8 7 3 21 19 1 15 14 18]

(vgl. Tabelle 5.2)

Im Folgenden wird der AMDF-Algorithmus mit Alg1, der patmat-Algorithmus (dieser Algorithmus wird in dem ursprünglichen Interpolationsprogramm verwendet und entspricht einer anderen Art von Differenzfunktion, vgl. [20] Formel (6)) mit Alg2 und der ZCR-Algorithmus mit Alg3 bezeichnet.

Auch bei den FSs gibt es Kürzungen: Es werden nur mehr zwei FSs mit je 2% Ausfälle der Länge 2-6ms verwendet (vgl. Tabelle 5.1). Die 1ms-Ausfälle werden nicht berücksichtigt, da diese leicht zu verschleiern sind. Ausfälle ab 7ms treten eher selten bis gar nicht auf und werden deshalb auch vernachlässigt.

Fehlerverteilung	
Länge in [ms]	Anzahl
2	9
3	6
4	5
5	4
6	3

Tabelle 5.1: VS3, Fehlerverteilung

Aus diesen Änderungen ergibt sich die Reihenfolge der Testbeispiele, wie sie in Tabelle 5.2 dargestellt ist.

SASQ - Testbeispiele					
Trial	Audiobeispiel	Fehlerszenario	Algorithmus	Original-Bezeichnung	Trial SASQ
1	1	1	1	AB1_DO1_AMDF6.wav	AB1_1
2	2	1	1	AB2_DO1_AMDF6.wav	AB2_2
3	3	1	1	AB3_DO1_AMDF6.wav	AB3_3
4	1	1	2	AB1_DO1_patmat.wav	AB1_4
5	2	1	2	AB2_DO1_patmat.wav	AB2_5
6	3	1	2	AB3_DO1_patmat.wav	AB3_6
7	1	1	3	AB1_DO1_zcr6.wav	AB1_7
8	2	1	3	AB2_DO1_zcr6.wav	AB2_8
9	3	1	3	AB3_DO1_zcr6.wav	AB3_9
10	1	2	1	AB1_DO3_AMDF6.wav	AB1_10
11	2	2	1	AB2_DO3_AMDF6.wav	AB2_11
12	3	2	1	AB3_DO3_AMDF6.wav	AB3_12
13	1	2	2	AB1_DO3_patmat.wav	AB1_13
14	2	2	2	AB2_DO3_patmat.wav	AB2_14
15	3	2	2	AB3_DO3_patmat.wav	AB3_15
16	1	2	3	AB1_DO3_zcr6.wav	AB1_16
17	2	2	3	AB2_DO3_zcr6.wav	AB2_17
18	3	2	3	AB3_DO3_zcr6.wav	AB3_18
19	1	-	-	AB1.wav	AB1
20	3	-	-	AB3.wav	AB3
21	1	1	-	AB1_DO_1_2pc_2_6ms_5s_v6.wav	AB1_DO
22	2	2	-	AB2_DO_3_2pc_2_6ms_5s_v6.wav	AB2_DO

Tabelle 5.2: VS3, Übersicht über die verwendeten Audiobeispiele

Für die folgenden Auswertungen werden die Bewertungen der Placebobeispiele vernachlässigt (nur Bewertungen von AB1-AB18), außer es wird dezidiert darauf hingewiesen (vgl. Abschnitt 5.2.2).

5.2 Auswertung der Ergebnisse

Das erste Diagramm, Abb. 5.1, zeigt die Bewertungen der VS3 der einzelnen Probanden. Beim Vergleichen der Bewertungen der einzelnen VPN lassen sich deutliche Unterschiede bei den Bewertungsintervallen erkennen.

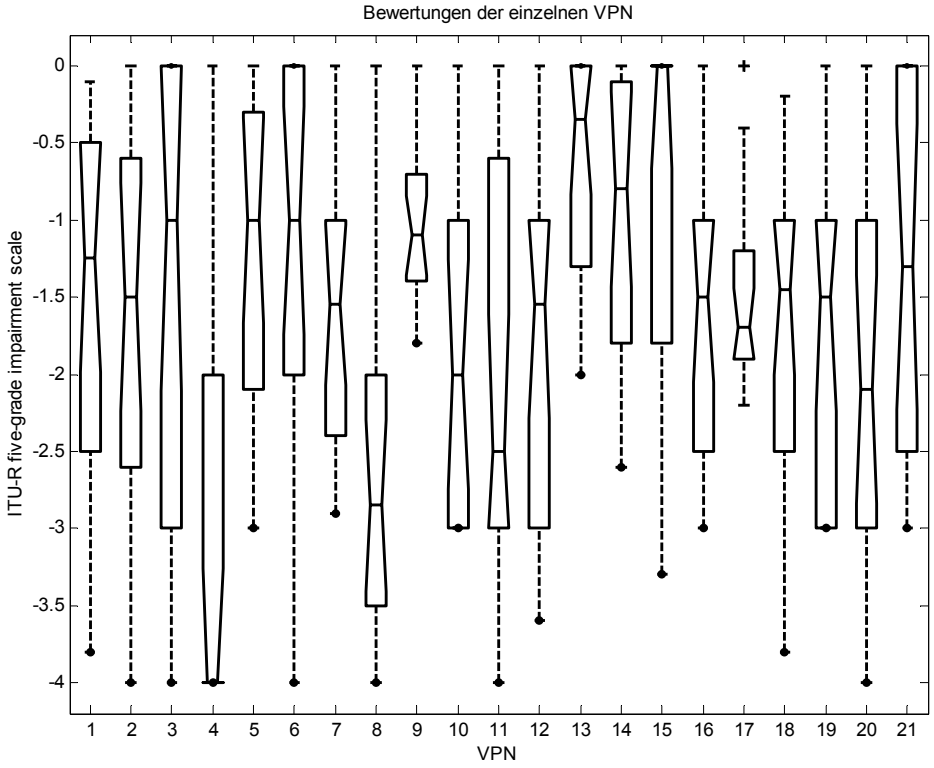

Abb. 5.1: VS3, Bewertung der einzelnen VPN

Die VP Nr.4 und Nr.15 sind zwei Extrema: $Q2_{Nr.4} = -4$ bzw. $Q2_{Nr.15} = 0$, d.h. 50% der Bewertungen werden bei Nr.4 mit „very annoying“ bewertet, während die Ergebnisse von Nr.15 50% Fehlbewertungen aufweisen. Der erste Gedanke lässt auf einen „expert“ und einen „non-expert listener“ schließen. Durch die Kontrolle der Namensliste ergibt sich, dass beide Bewertungen aus der gleichen Personengruppe stammen. Ein weiteres Beispiel ist VP Nr.8. Aus der Graphik lässt sich erkennen, dass diese VP eine sehr kritische Bewertung abgegeben hat. Auch hier ist die Vermutung falsch, die Bewertung eines „expert listeners“ vor sich zu haben. VP Nr.8 fällt in die Kategorie „Personen ohne musikalischer Erfahrung“.

Einen hohen Anteil an Fehlbewertungen ($Q1 = 0$, das entspricht 25%) haben VP Nr.3, Nr.6, Nr.13, Nr.14, Nr.15 und Nr.21. Zwei VPN (Nr.9 und Nr.17) nutzen einen extrem eingeschränkten Skalenbereich.

Aus diesen Ergebnissen lässt sich schließen, dass die VPN trotz vorgegebener Skala sehr individuell bewerten. Damit dies nicht so stark ins Gewicht fällt, ist es eine Überlegung wert, in Zukunft eine nachträgliche Anpassung der SDGs durchzuführen.

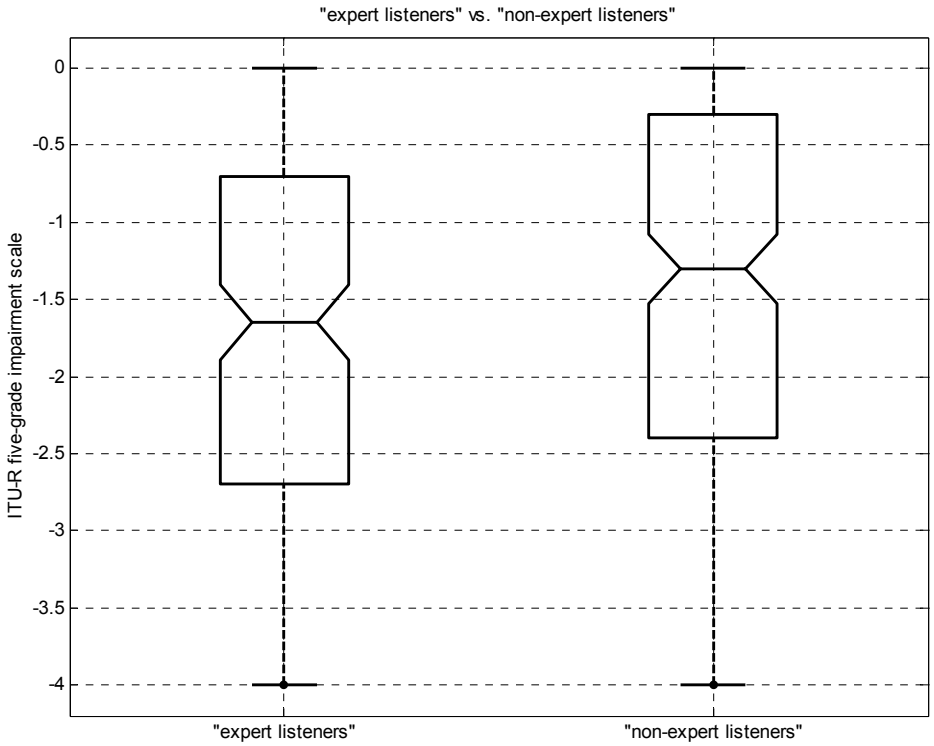

Abb. 5.2: VS3, Gegenüberstellung der Bewertungen von „expert“ und „non-expert listeners“

Über alle Bewertungen lässt sich die Tendenz einer besseren Bewertung durch „non-expert listeners“ erkennen (vgl. Abb. 5.2). Da sich die beiden CIe überschneiden, kann man die Differenz der Medianwerte ($Q2_{diff} \approx 0.3$) nicht als eindeutiges Qualitätskriterium verwenden. Die linke Box entspricht im Prinzip der um 0.3 nach unten verschobenen rechten Box (beide Grenzwerte (UG und OG) stimmen überein, nur die Quartile unterscheiden sich jeweils um ungefähr 0.3).

5.2.1 Auswirkung der FSs

Signifikante Unterschiede zwischen den beiden FSs ergeben sich nur beim Jazzbeispiel (vgl. Bewertungen der FSs der AMDF und der ZCR). Bei den anderen Boxplots überschneiden sich jeweils die CIe (vgl. Abb. 7.5 und Abb. 7.6). Für die nächsten Betrachtungen werden die Daten der FSs zusammengefasst.

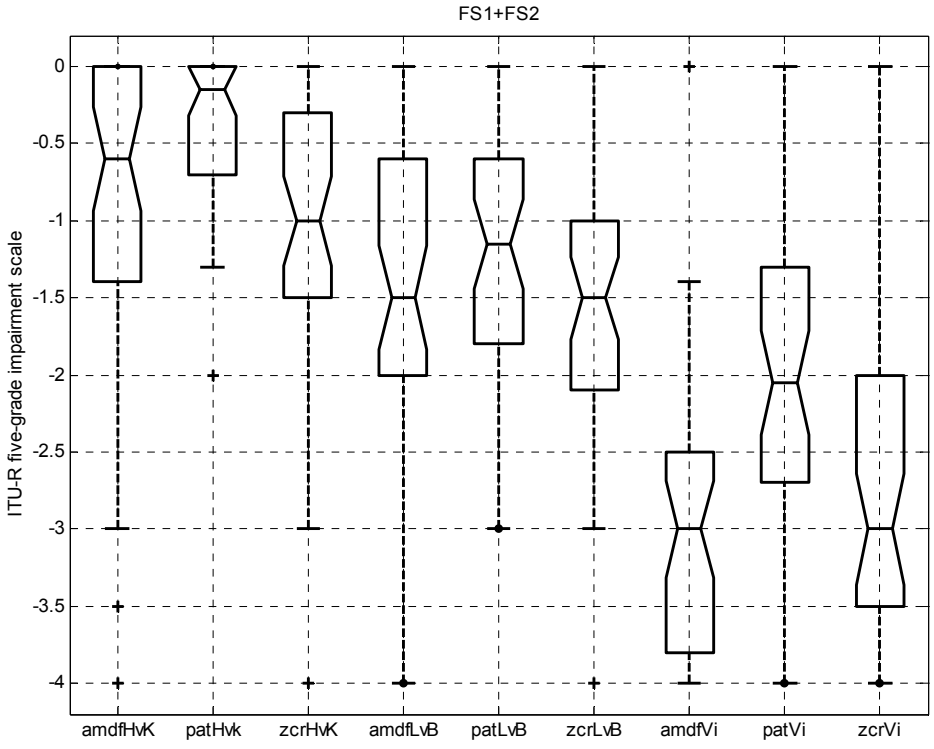

Abb. 5.3 VS3, Gegenüberstellung der Algorithmen und der Audiobeispiele (FS1+FS2)

Das Diagramm in Abb. 5.3 zeigt die Bewertungen der Audiobeispiele in Abhängigkeit der Algorithmen. Man erkennt deutlich den Qualitätsverlust aller drei Algorithmen, angefangen beim Jazzbeispiel gefolgt vom Orchesterbeispiel und abschließend das Violinbeispiel. Da sich die CIe der Boxplots innerhalb eines Algorithmus' nicht überschneiden wird diese Aussage bestätigt: AB1 wird von allen Algorithmen am besten bewertet, dann AB2 und zuletzt AB3. Der Alg2 berechnet AB1 und AB3 wesentlich besser als Alg1 bzw. Alg3. Bis auf das AB1 ($Q2_{diff,amdfHvH,zcrHvK} = 0.4$, die CIe überschneiden sich jedoch) können Alg1 und Alg3 als gleichwertig angesehen werden. Die CIe der Bewertungen von AB2 überlappen sich für jeden Algorithmus, weshalb hier die Algorithmen als gleichwertig angesehen werden können.

5.2.2 Fehlbewertungen

Eine Auflistung, wann eine Bewertung als Fehlbewertung gewertet wird, findet man unter Abschnitt 4.3.2.2.

Die Gesamtanzahl der Bewertungen der VS3 beträgt: $22\ Trials \cdot 21\ VPN = \underline{\underline{462\ Bewertungen}}$;

ohne Placebobeispiele ergeben sich 378 Bewertungen, 61 (16.14%) davon sind Fehlbewertungen.

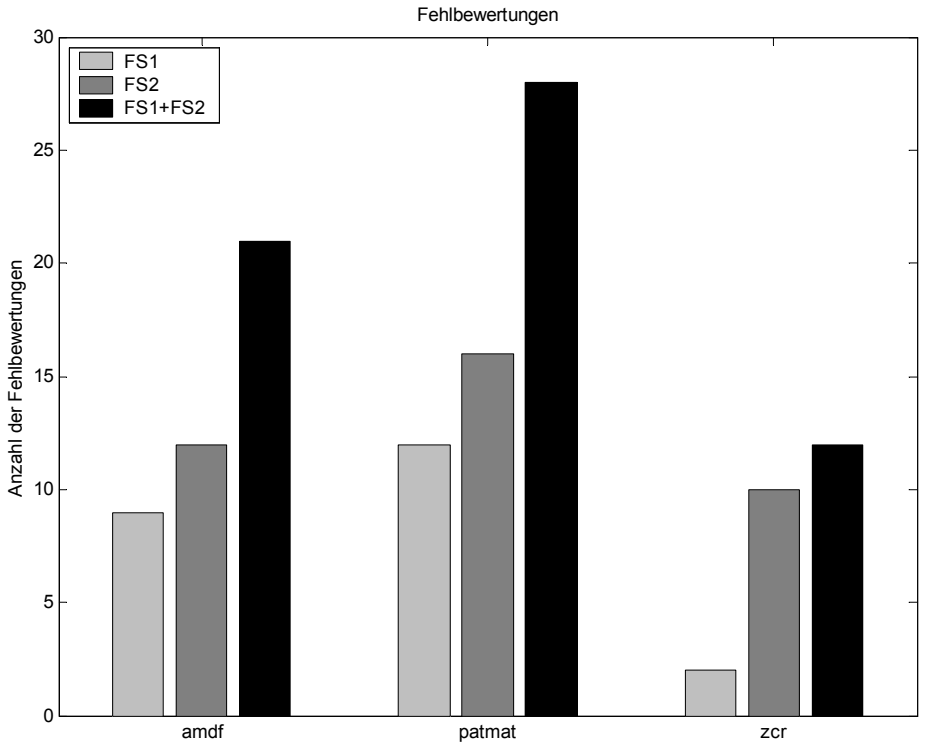

Abb. 5.4: VS3, Fehlbewertungen: FS1, FS2 und FS1+FS2

Fehlbewertungen - VS3					
Audiobeispiel	[Anzahl]	[%]	Algorithmus	[Anzahl]	[%]
Jazz	40	65.57	Alg1	21	34.43
Orchester	15	24.59	Alg2	28	45.90
Violine	06	9.84	Alg3	12	19.67
Summe: 61	61	100	Summe: 61	61	100

Tabelle 5.3: VS3, Fehlbewertungen (ohne Placebobeispiele)

Das Diagramm Abb. 5.4 stellt die Fehlbewertungen der einzelnen Algorithmen bzw. der FSs dar. Daraus lässt sich ablesen, dass Alg2 45.90%, Alg1 34.43% und Alg3 19.67% Fehlbewertungen (über die drei Audiobeispiele gerechnet) aufweisen, d.h. die fehlerverschleierten Audiobeispiele von Alg2 sind schwerer zu erkennen als jene von Alg1

bzw. Alg3. Bei einem Vergleich zwischen den Fehlbewertungen der VS1 bzw. VS2 mit jenen von VS3 erkennt man eine ungefähre Übereinstimmung der Fehlbewertungen des Jazzbeispiels. Die Anzahl der Fehlbewertungen des Orchester- und des Violinbeispiels sind jedoch gestiegen. Laut diesen Daten eignen sich die untersuchten Erweiterungsalgorithmen für die letztgenannten Audiobeispiele besser, als der ursprünglich verwendete Algorithmus.
Man sollte bei diesen Werten und Aussagen auch immer die Gesamtanzahl der Trials der VS3 (sie ist um ca. ein drittel kleiner) im Vergleich zu den anderen beiden VDN im Kopf behalten!

Abb. 5.5 zeigt eine Gegenüberstellung der Bewertungen der Placebobeispiele der VS3. Sieben Probanden (drei „expert listeners“, einer hat alle beide nicht erkannt, und drei „non-expert listeners“) haben eines, oder beide Originalbeispiele nicht erkannt und dementsprechend bewertet.

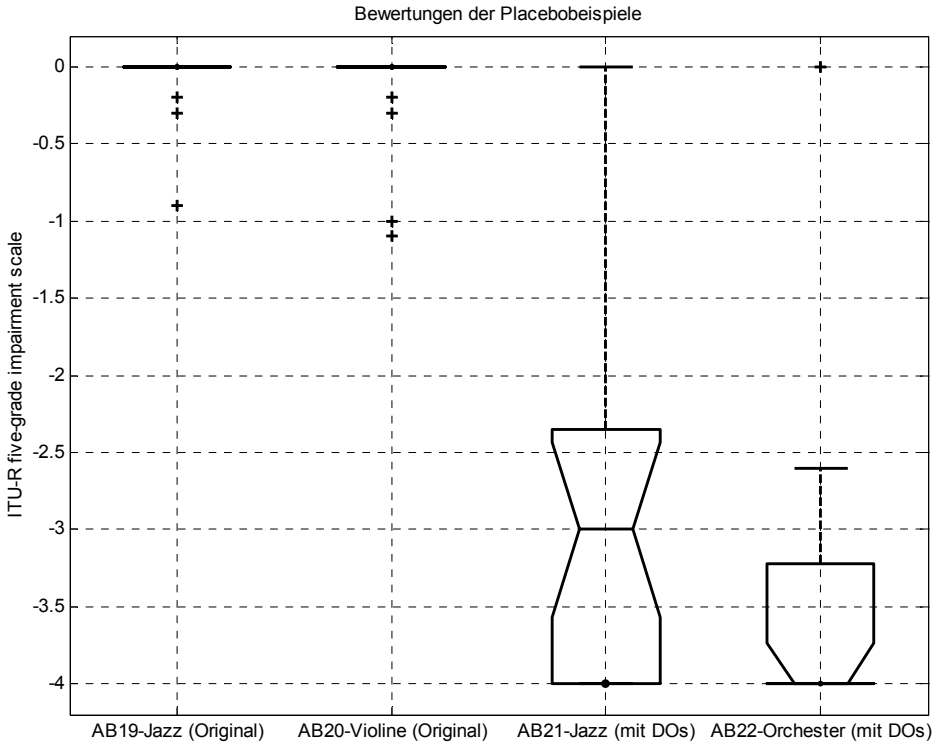

Abb. 5.5: VS3, Bewertungen der Placebobeispiele

Auffallend ist der große Wertebereich des dritten Boxplots. Der UG ist für dieses Beispiel eindeutig zu positiv. Dieser Wert kommt daher, dass zwei VPN das AB21 (vgl. Tabelle 5.2) mit Null bewertet haben (!). Bei diesen Bewertungen dürfte es sich um eine Verwechslung der Skalen von Seiten der jeweiligen VP handeln (d.h. „post-screening“ wäre hier sinnvoll).
Aus der Graphik ist auch ersichtlich, dass die unverschleierten DOs im Jazzbeispiel nicht so

störend empfunden werden wie beim Orchesterbeispiel. Durch Vergleichen der Boxplots von fehlerverschleierten Jazzbeispielen mit jenen des Jazzbeispiels mit unverschleierten DOs, erkennt man aber, dass der Bewertungsbereich von fehlerverschleierten Jazzbeispielen immer deutlich über dem Bereich des Placebobeispiels liegt. Der Medianwert des Orchesterbeispiels entsprich mit -4 jenem eines Signals mit unverschleierten DOs.

5.2.3 Bewertungsstatistik

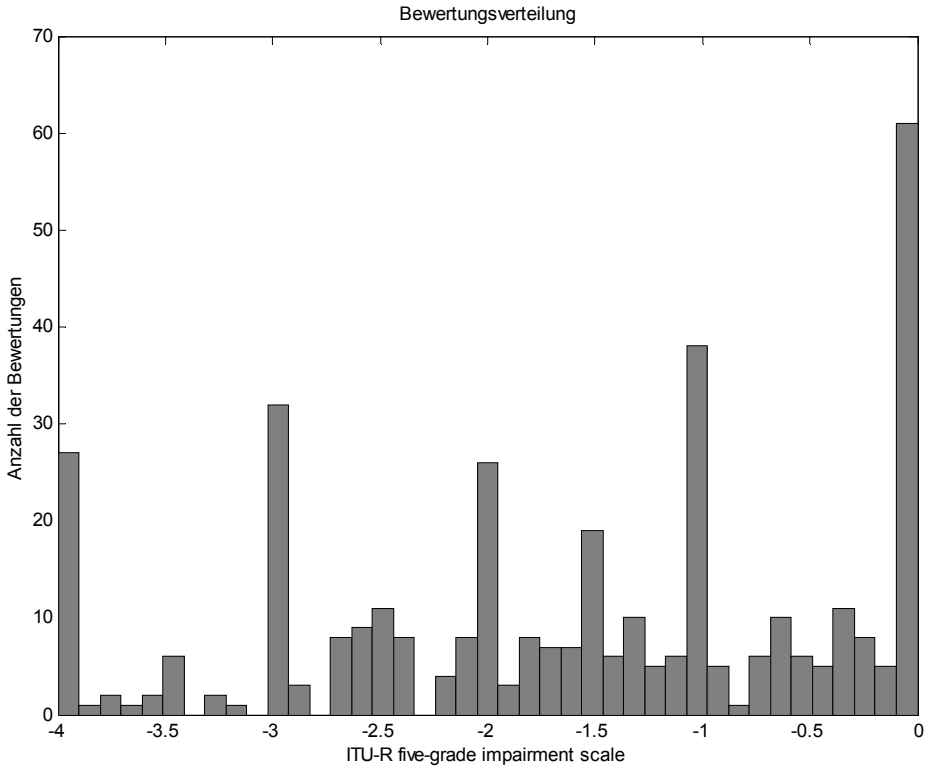

Abb. 5.6: VS3, Verteilung der subjektiven Bewertungen aller VPN

Die Graphik Abb. 5.6 veranschaulicht die Häufigkeitsverteilung der subjektiven Bewertungen aller VPN bezogen auf die ITU-Skala. Deutlich erkennt man eine starke Fokussierung auf ganze (negative) Zahlen bzw. auf 0.5er Zwischenschritten (vlg. Abb. 4.18 und Abb. 4.31). Im Intervall [0,-3] ist die Anzahl der Bewertungen fast gleich verteilt. Ähnlich den Ergebnissen der VS1 und VS2 liegen bei VS3 mit 66.93% (zwei Drittel aller Werte!) die meisten Bewertungen im Intervall $[0,-2]$.

Die Bewertungsverteilung kann immer noch als eine rechtssteile Verteilung angesehen werden, jedoch mit einem weniger steilen Anstieg und einem kleineren Maximalwert (daraus folgt eine wesentlich breitere „Spitze“).

Intervall	VS3	
	[Anzahl]	[%]
[0,-2]	253	66.93
]-2,-4]	125	33.07
Summe: 378 (100%)	**378**	**100**

Tabelle 5.4: VS3, Auflistung der Anzahl der Bewertungen

In Summe ergibt sich wieder eine positive Bewertung der Algorithmen, da die Mehrheit der VPN die Artefakte in den Versuchsbeispielen als „slightly annoying“ bzw. „perceptible, but not annoying“ beurteilt haben.

5.2.4 Zusammenfassung

Betrachtet man Abb. 5.7, erkennt man einige Unterschiede zwischen den Bewertungen der einzelnen Algorithmen.

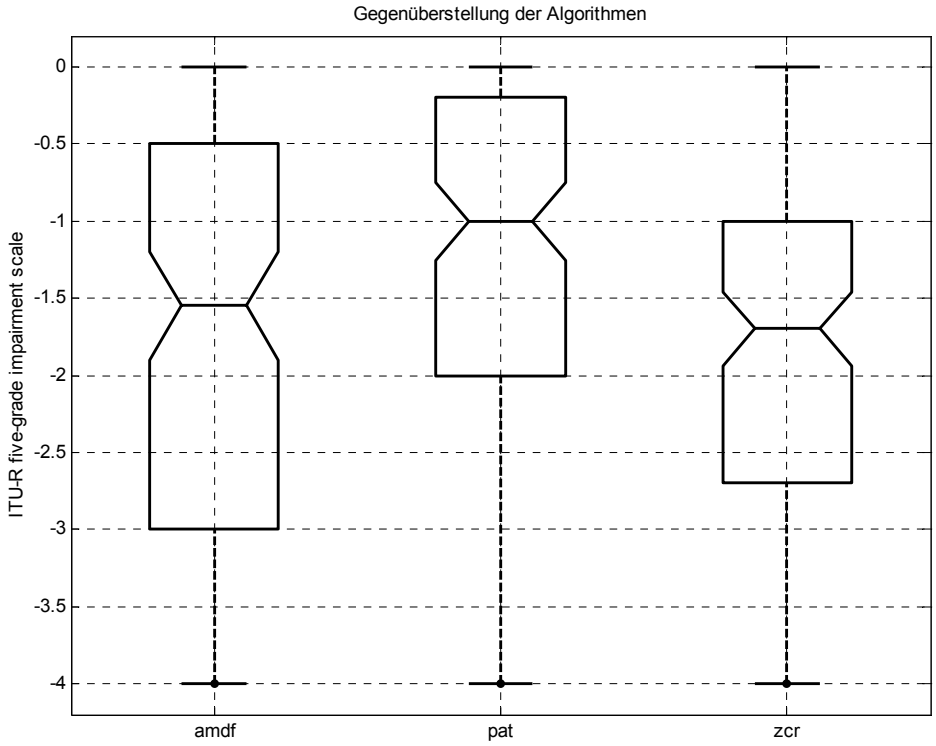

Abb. 5.7: VS3, Gegenüberstellung der Algorithmen (über alle drei Audiobeispiele)

Die CIe von Alg1 und Alg2 überschneiden sich gerade noch, weshalb die große Differenz der Medianwerte $Q2_{diff,Alg1,Alg2} \approx 0.6$ alleine nicht aussagekräftig genug ist. Durch Vergleichen ihrer $Q1$ bzw. $Q3$ lässt sich auf eine qualitativ bessere Bewertung von Alg2 schließen.

Die Boxplots von Alg1 und Alg3 haben zwei fast idente CIe ($Q2_{diff,Alg1,Alg3} = 0.1$), eine große Differenz $Q1_{diff,Alg1,Alg3} = 0.5$ und einen kleinen Unterschied bei $Q3_{diff,Alg1,Alg3} = 0.3$.

Aus diesen Ergebnissen geht Alg2 als der effizienteste hervor. Die beiden anderen Algorithmen können hingegen als gleichwertig angesehen werden; der ZCR-Algorithmus arbeitet jedoch wesentlich recheneffizienter und schneller als der AMDF-Algorithmus!

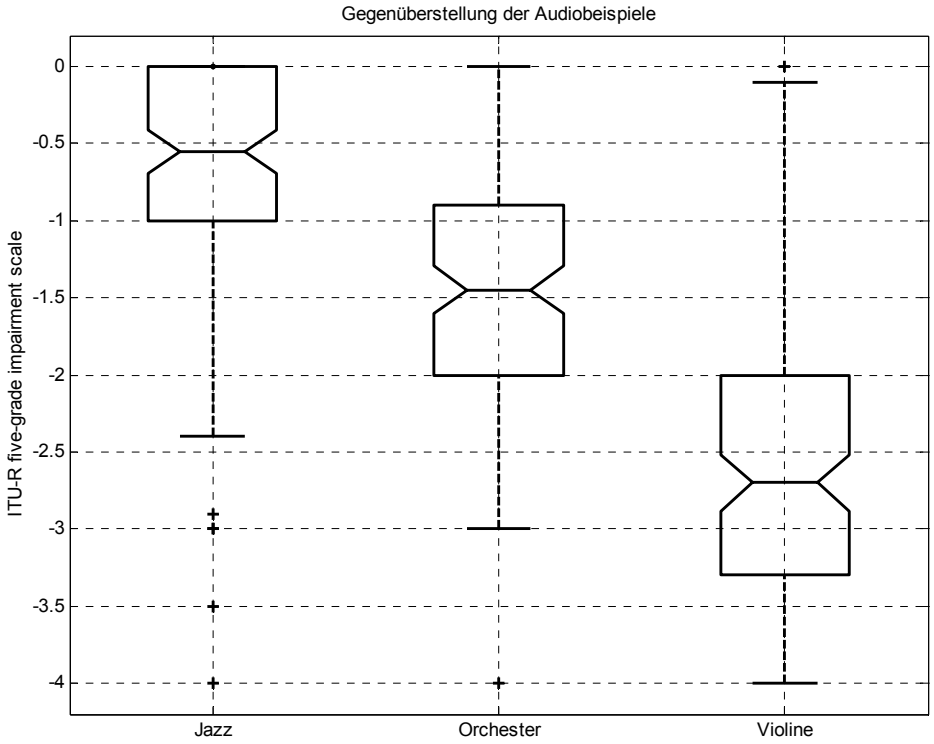

Abb. 5.8: VS3, Gegenüberstellung der Audiobeispiele (über die drei Algorithmen)

Im Gegensatz zur Auswertung der einzelnen Algorithmen kann man bei den Bewertungen der Audiobeispiele signifikante Unterschiede erkennen (vgl. Abb. 5.8).

Beim Jazzbeispiel treten mit 25% ($Q1_{Jazz} = 0$) am meisten Fehlbewertungen auf. Da 75% aller Bewertungen im Intervall $[0,-1]$ („imperceptible“, „perceptible, but not annoying“) liegen, ist dies ein weiterer Indikator für eine sehr gute Bewertung dieses Beispiels. Die Bewertungen des Jazzbeispiels liegen in einem eher kleinen Skalenbereich $[0,-2.4]$ (bis auf vier Ausreißer).

Beim Orchesterbeispiel treten auch noch häufiger Fehlbewertungen auf, 75% der Bewertungen liegen jedoch im Intervall $[0,-2]$.

Am besten wird die Skala von den Ergebnissen des Violinebeispiels ausgenutzt $[-0.1,-4]$.

Bis auf eine einzige Fehlbewertung (ein Ausreißer bei Null) sind in dieser Bewertung alle fehlerverschleierten Signale detektiert worden.
Die Differenzen der Medianwerte $Q2_{diff,Alg1,Alg2} \approx 0.9$, $Q2_{diff,Alg1,Alg3} \approx 2.1$ und $Q2_{diff,Alg2,Alg3} \approx 1.1$ sind beträchtlich. Da sich keines der CIe mit einem anderen überlappt, ist dies eine zusätzliche Bestätigung für den Bewertungs- bzw. Qualitätsunterschied.

Das Ranking der Audiobeispiele nach ihren Bewertungen erfolgt wie schon in Abb. 5.8 dargestellt: Das Jazzbeispiel wird am besten bewertet, gefolgt vom Orchesterbeispiel und zum Schluss mit der schlechtesten Bewertung das Violinebeispiel.

6 Zusammenfassung / Ausblick

Resümee

Für den ZCR-Algorithmus werden in dieser Arbeit zwei Ansätze betrachtet: Zählen der Nulldurchgänge (inklusiv ihrer Position) bzw. Berechnungen diverser Energieanteile. Anschließend werden sie zusammengefügt und getestet.
Die Implementierung bzw. Auswertung der Qualität des ZCR-Algorithmus’ soll zeigen, dass man auch mit einer relativ einfachen Methode ein brauchbares Ergebnis für einen PM-Algorithmus bekommen kann. Nachdem diese Methode funktioniert, ist es eine Überlegung wert, den Algorithmus zu verbessern.
Entweder verwendet man die vorhandenen Ansätze und erweitert diese, oder man überlegt sich effizientere Methoden für einen zukünftigen ZCR-Algorithmus.

Die Auswertungen der VS1 und VS2 zeigen, dass die Einführungsphase einen wesentlichen Einfluss auf die Ergebnisse der Versuche hat. Dementsprechend umfangreich soll sie für zukünftige Hörversuche gestaltet werden, um die Probanden bestmöglich auf den Versuch vorzubereiten. Weiters ist aus den Ergebnissen eine Tendenz der schlechteren Bewertung des VD2 erkennbar. Diese kann auf Konzentrationsschwierigkeiten der Probanden aufgrund der langen Versuchsdauer zurückgeführt werden.
Die VS3 wird entsprechend der Erkenntnisse der vorherigen VSN gestaltet: es werden nur Probanden, die VS1 oder VS2 absolviert haben, für diese VS herangezogen. Dadurch kann auf die Einführungsphase verzichtet werden (VS1 bzw. VS2 gilt in diesen Fällen als Einführungsphase). Zweitens wird die Dauer des Versuchs wesentlich herabgesetzt (u.a. wird nur ein VD durchgeführt). Mit Hilfe dieser Modifikationen bekommt man aussagekräftigere Ergebnisse.
Eine weitere Verbesserung wäre ein „post-screening“ der VPN gewesen (bei drei bis vier Fällen). Da in Summe wenig Probanden vorhanden sind, wird bei den Auswertungen auf dieses Auswahlverfahren verzichtet.

„Short-term, long-term prediction“ – STLTP

Eine mögliche Verbesserung des Extrapolationsalgorithmus’ ist die STLTP. Diese besteht, wie aus dem Namen schon ersichtlich, aus der Kombination von zwei Prädiktionen (Kurz- und Langzeitkorrelation).

Bei der Kurzzeitprädiktion werden die Prädiktionskoeffizienten mit Hilfe der letzten P-Samples vor dem Extrapolationszeitpunkt berechnet. Für die Berechnung der Koeffizienten der Langzeitprädiktion wird der um die Grundperiode T_0 in die Signalvergangenheit verschobene Signalbereich $[T_0 - Q, T_0 + Q]$ verwendet (vgl. Abb. 6.1), wobei die Prädiktionsordnung nicht Q, sondern $2 \cdot Q + 1$ ist. Anschließend wird das Ergebnis der Langzeitprädiktion unter Berücksichtigung des Prädiktionsfehlers zum Ergebnis der Kurzzeitprädiktion addiert.

(a) „Short-term prediction“: berechnet die Korrelation jedes Samples mit den P vergangenen Samples: $x(n-1), \ldots, x(n-P)$ (vgl. Abb. 6.1).

(b) „Long-term prediction“: darunter versteht man die Korrelation eines Samples $x(n)$ mit $2Q+1$ ähnlichen, um eine Grundperiode T in die Signalvergangenheit verschobenen Samples: $x(n-T+Q), \ldots, x(n-T-Q)$ (vgl. Abb. 6.1).

Die „short-term“ Korrelation $\hat{x}(n)$ eines Signals kann mit Hilfe der linearen Prädiktion ausgedrückt werden (vgl. Formel (6.1)).

$$\hat{x}(n) = -\sum_{k=1}^{p} a(k)x(n-k) \tag{6.1}$$

Ihr Prädiktionsfehler wird mit folgender Gleichung berechnet:

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{k=1}^{p} a(k)x(n-k) \tag{6.2}$$

Aus Formel (6.2) lässt sich eine Datenmatrix erstellen (u.a. für die Erklärung der AK in Abschnitt 2.4):

$$\begin{pmatrix} e(1) \\ \vdots \\ e(p+1) \\ \vdots \\ e(N-p) \\ \vdots \\ e(N) \\ \vdots \\ e(N+p) \end{pmatrix} = \underbrace{\begin{pmatrix} x(1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ x(p+1) & & x(1) \\ \vdots & \ddots & \vdots \\ x(N-p) & & x(p+1) \\ \vdots & \ddots & \vdots \\ x(N) & & x(N-p) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x(N) \end{pmatrix}}_{X_p} \begin{pmatrix} 1 \\ a(1) \\ \vdots \\ a(p) \end{pmatrix} \tag{6.3}$$

Der Fehler der „long-term prediction“ $e'(n)$ wird mit Hilfe der Grundperiode berechnet. Die Koeffizienten $p(k)$ der „long-term prediction“ haben die Ordnung $2Q+1$.

$$e'(n) = \sum_{k=-Q}^{Q} p(k)e(n-T-k) \tag{6.4}$$

$$\varepsilon(n) = e(n) - e'(n) = e(n) - \sum_{k=-Q}^{Q} p(k)e(n-T-k) \tag{6.5}$$

Bei einer Echtzeitanwendung der STLTP gilt: $e(n) = 0$ (vgl. Formel (6.2)).

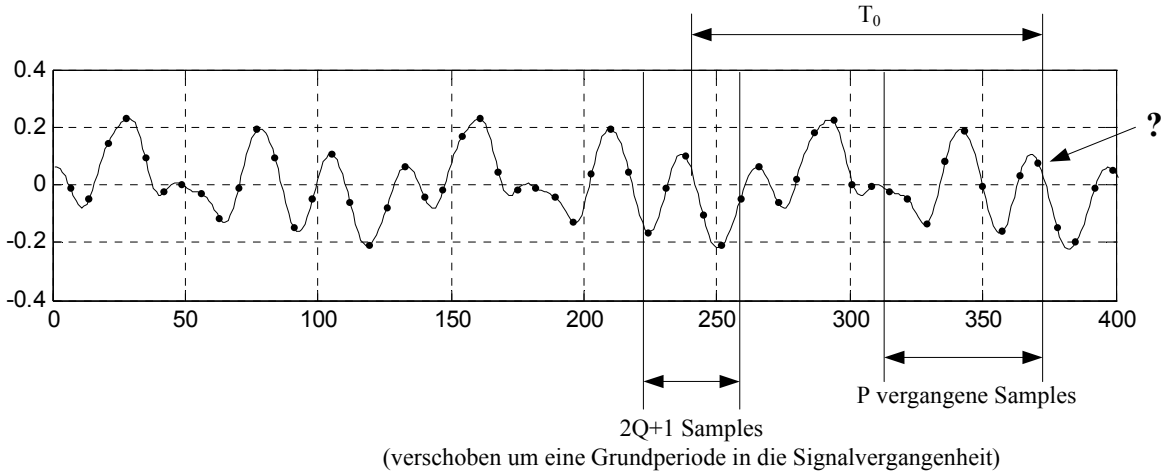

Abb. 6.1: Graphische Darstellung der „short-term, long-term prediction“

Eine Vereinfachung der separaten Berechnung der beiden Korrelationen bietet die Kombination der „short“ und „long-term prediction“:

$$\hat{x}(n) = \underbrace{\sum_{k=1}^{P} a(k)x(n-k)}_{STP} + \underbrace{\sum_{k=-Q}^{Q} p(k)x(n-k-T) + \varepsilon(n)}_{LTP} \tag{6.6}$$

CD zur Diplomarbeit

Auf der letzten Seite (Einbandinnenseite) befindet sich eine CD. Darauf sind alle verwendeten „papers“, die MATLAB®-Programme (inklusiv Hörversuch), die „dropout“-Statistiken und die verwendeten Audiobeispiele (originale und verschleierte) enthalten. Die Diplomarbeit ist zusätzlich als PDF-Datei vorhanden.

7 Anhang

Die folgenden Diagramme dienen als Ergänzungen für die Auswertungen der drei VSN. In den Graphiken Abb. 7.1 bis Abb. 7.4 sind die Bewertungen der beiden Versuchsdurchläufe eines Algorithmus' gegenübergestellt. Abb. 7.5 und Abb. 7.6 zeigen die Ergebnisse der zwei verwendeten FSs.

Verwendete Abkürzungen in den folgenden Graphiken:

bxvy … „b“ bedeutet Beispiel, „x“ steht für die entsprechende Nummer (1…6); „v“ steht für VD und „y“ entspricht wiederum der Nummer (1,2).

pat … steht für den patmat-Algorithmus

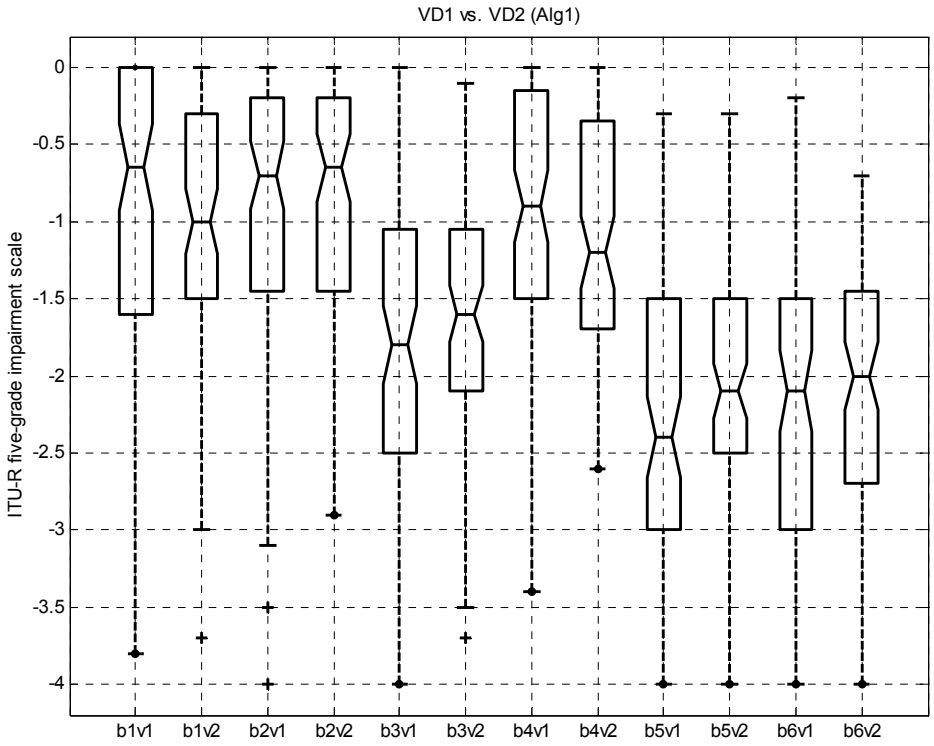

Abb. 7.1: VS1, Gegenüberstellung der Bewertungen des VD1 und VD2 (Alg1)

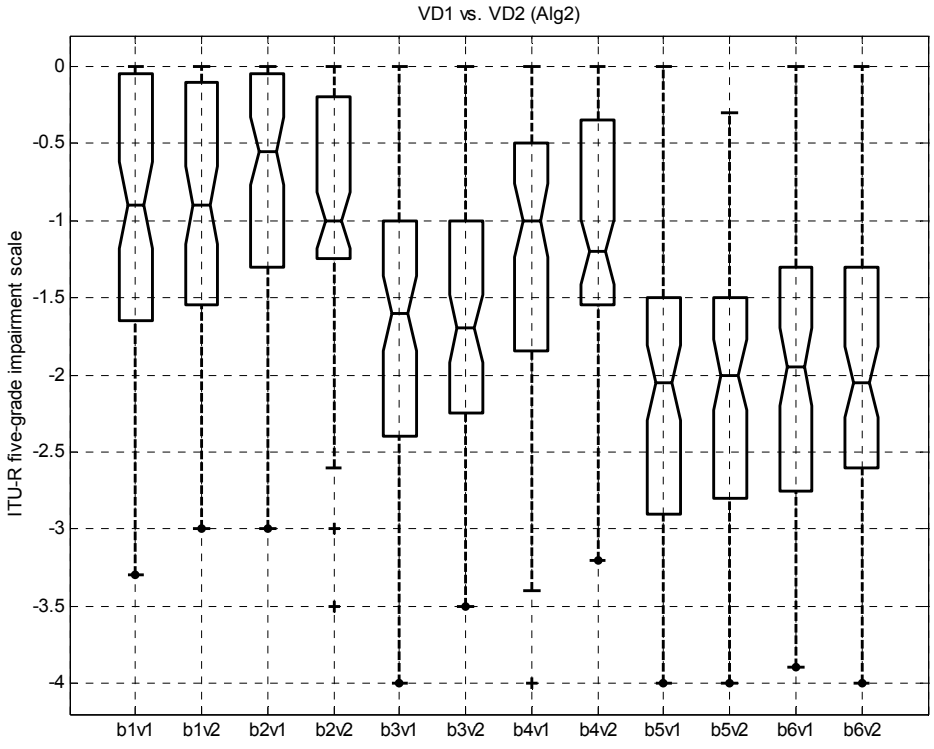

Abb. 7.2: VS1, Gegenüberstellung der Bewertungen des VD1 und VD2 (Alg2)

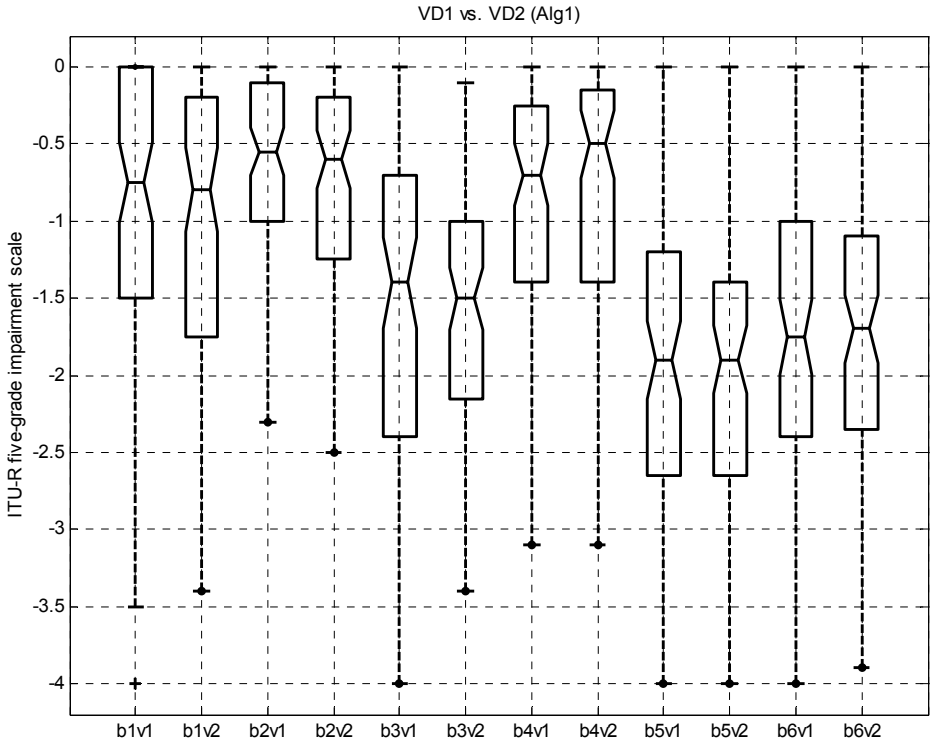

Abb. 7.3: VS2, Gegenüberstellung der Bewertungen des VD1 und VD2 (Alg1)

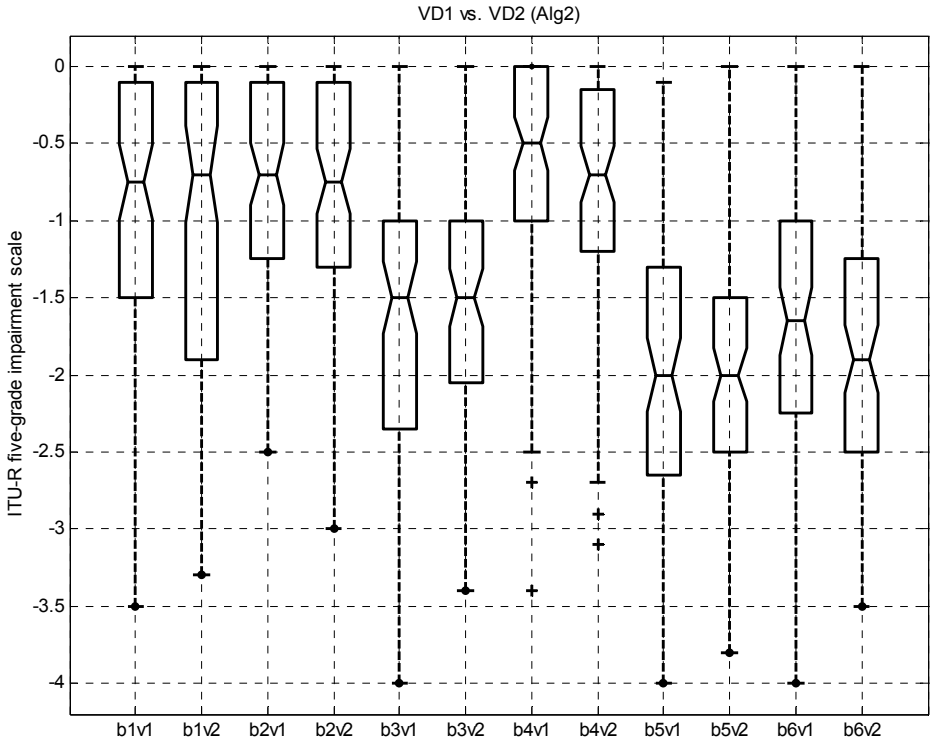

Abb. 7.4: VS2, Gegenüberstellung der Bewertungen des VD1 und VD2 (Alg2)

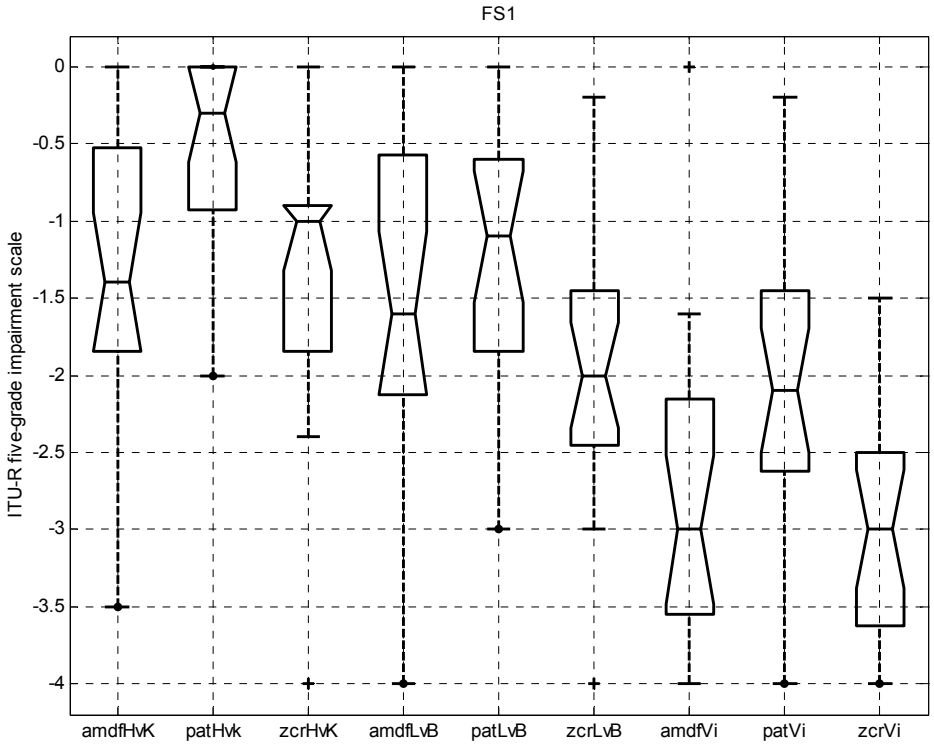

Abb. 7.5: VS3, Gegenüberstellung der Algorithmen und der Audiobeispiele, FS1

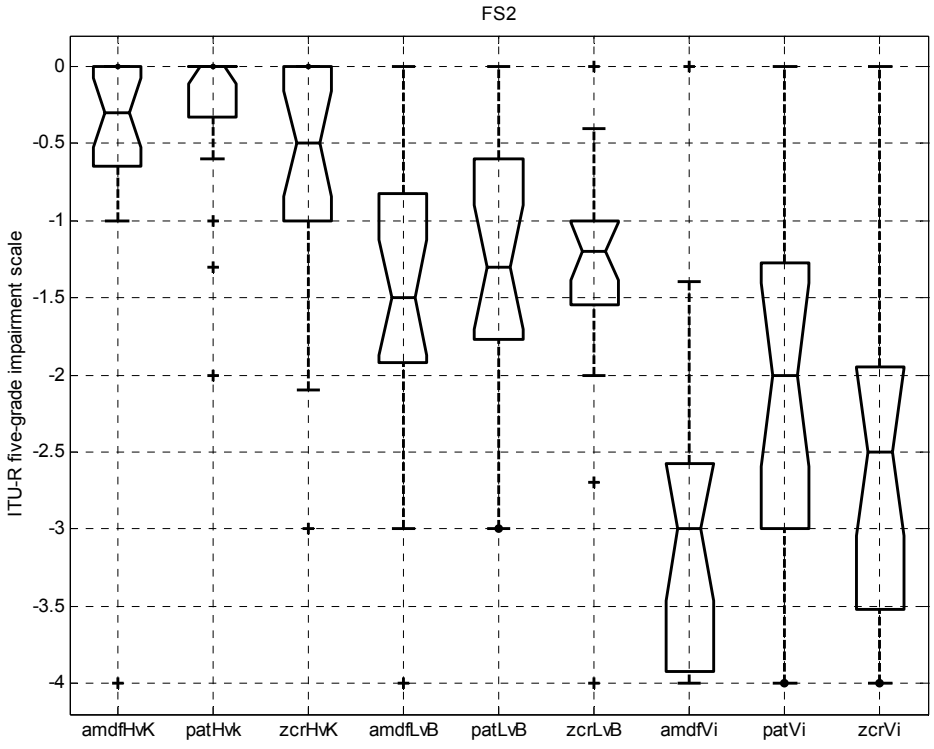

Abb. 7.6: VS3, Gegenüberstellung der Algorithmen und der Audiobeispiele, FS2

Literaturverzeichnis

[01] Recommendation ITU-R BS.562-3, (1978-1982-1986-1990), „Subjective Assessmet Of Sound Quality“

[02] Recommendation ITU-R BS.1116-1, (1994-1997), „Methods For The Subjective Assessmet Of Small Impairtments In Audio Systems Including Multichannel Sound Systems“

[03] Recommendation ITU-R BS.1284-1, (1997-2003), „General Methods For The Subjective Assessmet Of Sound Quality“

[04] Recommendation ITU-R BS.1387-1, (1998-2001), „Method Of Objective Measurements Of Perceived Audio Quality“

[05] Recommendation ITU-R BS.1523-1, (2001-2003), „Method For The Subjective Assessmet Of Intermediate Quality Level Of Coding Systems“

[06] Ilona Papousek, „Psychologische Statistik“ Handbuch (2002), Druck: Servicebetrieb der ÖH-Uni Graz, Sriptum Nr.: 445

[07] Bortz, Döring, „Forschungsmethoden und Evaluierung für Human- und Sozialwissenschaftler“, 3. Auflage, Springer, 2002, ISBN 3-540-41940-3

[08] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, B. Feiten, „PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality“, Journal of the Audio Engineering Society, Audio Acoustics Applications, Vol. 48, Nr. 1/2/3, January/February/March 2000

[09] T. Thiede, W. C. Treurniet, R. Bitto, T. Sporer, K. Brandenburg, C. Schmidmer, M. Keyhl, J. G. Beerends, C. Colomes, G. Stoll, B. Feiten, „PEAQ – der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität“, 1998

[10] OPERA Manual 3.5, Version 20.12.2002

[11] M. Karjalainen, „A New Auditory Model fort he Evaluation of Sound Quality of Audio Systems“, in Proc. ICASSP (Tampa, FL, März 1985), pp. 608-611

[12] P. Vary, U. Heute, W. Hess, Informationstechnik – „Digitale Sprachsignal-verarbeitung“, B. G. Teubner Stuttgart, 1998, ISBN 3-519-06165-1

[13] S. Lawrence Marple Jr., „Digital Spectral Analysis“ – With Applications, Prentice-Hall Signal Processing Series, 1987, ISBN 0-13-214149-3 025

[14] A. V. Oppenheim, R. W. Schafer, „Zeitdiskrete Siganlverarbeitung“, 3. durchges. Auflage, R. Oldenbourg Verlag München Wien, 1999, ISBN 3-486-24145-1

[15] Saeed V. Vaseghi, „Signal Processing and Digital Noise Reduction“, Queen’s University of Belfast, UK, 1996, ISBN Wiley 0-471-958751, ISBN Teubner 3-519-06451-0

[16] Alain de Cheveigne, Hideki Kawahara, „YIN, a fundamental frequency estimator for speech and music“, 9.January 2002, 0001-4966/2002/11(4)/1917/14

[17] Tetsuya Shimamura, Hajime Kobayashi, „Weighted Autocorrelation for Pitch Extraction of Noisy Speech“, IEEE transactions on speech and audio processing, Vol. 9, No. 7, October 2001, 1063-6676(01)08234-7

[18] Goangshiuan S. Ying, Leah H. Jamieson, Carl D. Michell, „A probabilistic approach to AMDF pitch detection“, Proceedings of the 1996 International Conference on Spoken Language Processing, Philadelphia, PA, Oct. 1996, pp. 1201-1204

[19] B. Fette, R. Gibson, E. Greenwood, „Windowing functions for the average magnitude difference function pitch extractor“, IEEE, 1980, CH1559-4/80/0000-0049

[20] David J. Goodman, Gordon B. Lockhart, Ondria J. Wasem, Wai-Choong Wong, „Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications“, IEEE transactions on acoustis, speech and signal processing, Vol. ASSP-34, No. 6, December 1986, 0096-3518/86/1200-1440

[21] David Gerhard, „Pitch Extraction and Fundamental Frequency: History and Current Techniques“, November 2003, ISBN 0828-3494

[22] Maciej Niedzwiecki, „Statistical Reconstruction Of Multivariate Time Series“, IEEE transactions on signal procession, Vol. 41, No. 1, January 1993, 1053-587X/93

[23] Maciej Niedzwiecki and Krzysztof Cisowski, „Smart Copying—A New Approach to Reconstruction of Audio Signals“, IEEE transactions on signal processing, Vol. 49, No. 10, October 2001, 1053-587X/01

[24] Ondria J. Wasem, David J. Goodmann, Charles A. Dvorak, Howard G. Page, „The Effect of Waveform Substitution on the Quality of PCM Packet Communications“, IEEE transactions on acoustics, speech and signal processing, Vol. 36, No. 3, March 1988, 0096-3518/88/0300-0342

[25] W. Hess, „Pitch Determination of Speech Signals“ – Algorithms and Devices, Springer-Verlag 1983, ISBN 3-540-11933-7

[26] Simon Haykin, „Adaptive filter theory”, Fourth Edition, Prentice Hall Information and System Scients Series, 2002, ISBN 0-13-048434-2