Auditive Repräsentation von visuellen Reizen in Processing

Finn Bayer (Matrikelnummer: 2364311)
Department Medientechnik
Fakultät Design, Medien und Information
Hochschule für angewandte Wissenschaften Hamburg
Email: finn.bayer@haw-hamburg.de

I. Abstract

Die sensorische Substitution ist zentraler Bestandteil der Hilfsmittel für Blinde. Dabei wird der Sehsinn durch eine auditive oder taktile Repräsentation des eigentlichen Reizes ersetzt. Untersuchungen im Bereich der visuell-auditiven Substitution werden seit mehr als 50 Jahren durchgeführt. Trotzdem ist eine praktische Nutzung der Hilfssysteme nur vereinzelt anzutreffen. Ein mögliches Problem der aktuellen Systeme stellt die festgelegte Auflösung dar. Eine Änderung dieser kann nur durch eine erneute ausgiebige Trainingsphase erzielt werden. In diesem Bericht wird ein Ansatz präsentiert, in dem dieser Umstand mithilfe der Hilbert-Kurve angegangen wird. Durch eine Umsortierung der Pixel anhand des Verlaufs der Hilbert-Kurve verliert eine Änderung der Auflösung ihren Einfluss auf die extrahierten Features des visuellen Reizes. Das Programm wurde in der Programmiersprache Processing realisiert und besteht neben der Neuanordnung der Pixel aus der Extraktion von Helligkeits- und Positionsinformationen aus den Pixeln zur Erstellung einer auditiven Repräsentation in Form eines Tones. Dieser besteht aus verschiedenen Frequenzen und Amplituden, die sich aus den extrahierten Informationen ergeben. Sowohl die Tests zur Funktionalität des Programms und Anwendbarkeit, als auch die anschließende Einordnung der Ergebnisse in den Kontext der aktuellen Forschung stehen noch aus.

II. Konzeptidee

Sensorische Substitution erlaubt es einen bestimmten Reiz von einem reizfremden Sinnesorgan interpretieren zu lassen. Einen der ersten Substitutionsansätze beschreiben Bach et al. [2], die versuchten visuelle in taktile Reize zu übertragen. Dies sollte dazu genutzt werden blinden Menschen eine weitere Möglichkeit der Orientierung bereit zu stellen. Im Laufe der Zeit wurden immer weitere Systeme entwickelt, deren Zweck in der Unterstützung von beeinträchtigten Personen liegen sollte. Dabei liegt der Fokus meist auf einer Substitution von visuellen Reizen [11] [5] [4]. Trotzdem werden diese Systeme auch fast 50 Jahre nach der Entwicklung der ersten Prototypen nur vereinzelt praktisch genutzt [9]. Im Kontext dieser Projektarbeit im Bereich Image- und Soundprocessing wird die Frage gestellt, welchen Stand die Forschung in dem Bereich der visuell-auditiven Substitution hat und wo mögliche Probleme liegen können. Einige dieser Probleme sollen dann in einer eigenen Implementierung angegangen werden.
Im Bereich der visuell-auditiven Substitution gibt es eine Vielzahl von verschiedenen Ansätzen. Hier werden exemplarisch drei verschiedene Forschungen aufgezeigt, an denen sich diese Arbeit weiter orientiert. Meijer [11] beschreibt in seinem Paper ein Low-Cost System, welches 64x64 Pixel große Graustufenbilder in Sound umwandelt. Dabei wird das Bild von links nach rechts gescannt und in einem Folge von hintereinander abgespielten Tönen übersetzt. Dadurch entsteht eine Abfolge von 64 Tönen pro Bild, bei denen die 64 Frequenzen eines Tons der vertikalen Position der Pixel entsprechen und die Lautstärke einer Frequenz kongruent zur Helligkeit des entsprechenden Pixels ist. Das Mapping der Frequenzen entsprechend der räumlichen Position der Pixel wird auch von Bernstein et al. [3] und Melara et al. [12] als zielführend beschrieben. Die Übersetzung der Helligkeit des Pixels in die Lautstärke der entsprechenden Frequenz beruhen auf den Ergebnissen von Marks [10], der einen starken Zusammenhang zwischen den entsprechenden Reizen festgestellt hat. Dieses von Meijer beschriebene System bildet die Grundlage für das sog. „The vOICe“ - System, welches heutzutage noch praktische Anwendung findet (s. [1]). Auch Cronly-Dillon et al. [6] nutzen den Ansatz von Meijer, jedoch versuchen sie die auditive Darstellung an der Komposition eines Musikstücks anzulehnen. Dabei werden z.B. die Frequenzen eines Pianos genutzt, um einen melodischen Klang herbeizuführen und damit die Simplizität des Ansatzes hinsichtlich der Anwendbarkeit für Blinde zu verbessern. Ein Problem dieser Ansätze besteht jedoch in der Auflösung der horizontalen Position in einem zeitlichen Kontext. So werden pro Bild mehrere Töne in dem Zeitraum von etwa einer Sekunde abgespielt. Dadurch wird die Bildfrequenz eingeschränkt und das Problem verstärkt sich weiter, wenn die Auflösung des Bildes vergrößert wird (je breiter das Bild, desto länger wird die Abfolge an Tönen). Deswegen nutzen Capelle et al. [5] ein

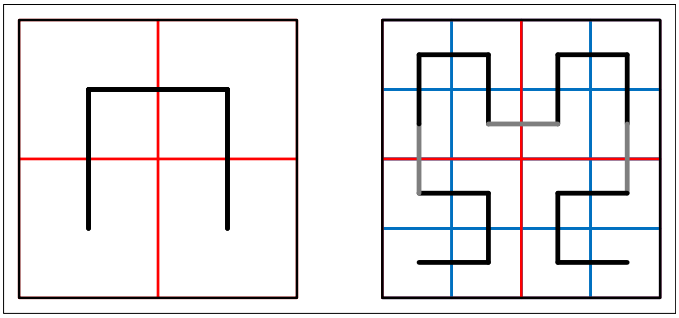

Abbildung 1: Die ersten beiden Ordnungen der Hilbert Kurve. Die erste Ordnung ist links dargestellt, die zweite Ordnung rechts.

Mapping der Pixel sowohl in vertikaler als auch in horizontaler Richtung. Dabei erhöht sich die Frequenz der Töne von links nach rechts und von unten nach oben. Dadurch ist die Anzahl der darzustellenden Bilder pro Sekunde frei wählbar, sodass trainierte Probanden eine höhere Framerate nutzen können. Hierbei kommt es jedoch zu dem Problem, dass die Auflösung des Bildes festgelegt ist. Bei einer Änderung der Auflösung ist ein neues Training vonnöten. Eine genaue Beschreibung dieses Problems gibt Sanderson [13] in seinem Video bei Minute 13:15. Sanderson zeigt in diesem Video ebenfalls eine Alternative für das Pixel-Frequenz Mapping auf. Dabei wird eine Hilbert-Kurve [8] verwendet, um die Pixel den einzelnen Frequenzen zuzuordnen. Dadurch wird ein bestimmter Pixelbereich immer dem gleichen Frequenzbereich zugeordnet, unabhängig von der Auflösung des Bildes. In Abbildung 1 wird dies am Beispiel von den ersten beiden Ordnungen der Hilbert-Kurve dargestellt. Jede Kante der Kurve symbolisiert dabei eine Frequenz. Für eine höhere Auflösung kann dann eine Hilbert-Kurve höherer Ordnung verwendet werden. Dieser Ansatz war Grundlage für das vorliegende Projekt.
Der Aufbau des Projekts ist in Abbildung 2 dargestellt. Als visueller Input soll zwischen verschiedenen Alternativen gewählt werden können. Zur Auswahl stehen dort festgelegte Trainingsbilder, das Bild einer angeschlossenen Webcam oder Bilder von der Festplatte (Schritt 1). Anschließend wird das eingehende Bild geladen und die Pixel entsprechend der genutzten Hilbert-Kurve sortiert (Schritt 2). In Schritt 3 wird das Mapping der visuellen Informationen auf die auditiven Reize vorgenommen. Dabei werden die folgenden Umwandlungen vorgenommen: Die Pixel werden entsprechend des Vorschlags von Sanderson [13] auf die Frequenzen gemappt. Dabei werden jedoch die Quadranten I, II und IV vertauscht, um den in Abbildung 3 dargestellten Frequenzverlauf zu erstellen, der den Ergebnissen der Forschungen von Bernstein et al. [3] und Melara et al. [12] entspricht. Die Frequenzen sind dabei, wie bei Cronly-Dillon [6], Frequenzen der gleichstufigen Stimmung, also die Töne eines Klaviers. Die Helligkeit der Pixel beschreibt die Lautstärke des Tons, genauso wie bei vorherigen Ansätzen. Weitere Informationen wie z.B. die Farben werden für dieses Projekt außer Acht gelassen. Die aus den Berechnungen gewonnenen Informationen werden in Schritt 4 genutzt, um den Sound zu generieren, der anschließend über die Lautsprecher ausgegeben wird (Schritt 5).

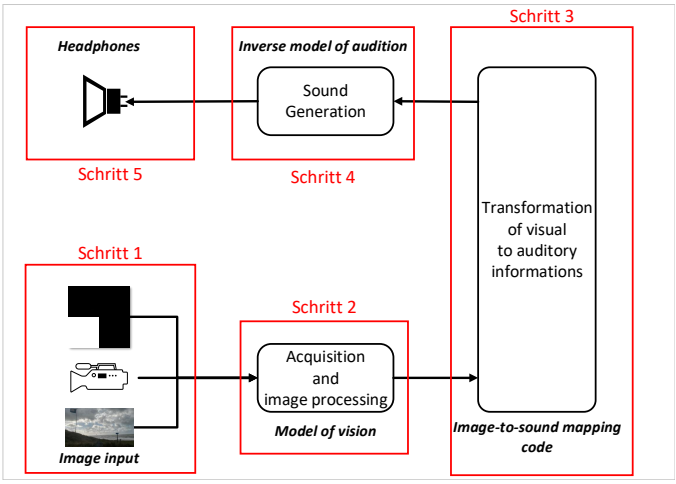

Abbildung 2: Visualisierung des Aufbaus basierend auf der Abbildung von Capelle et al. [5, Abb. 5]

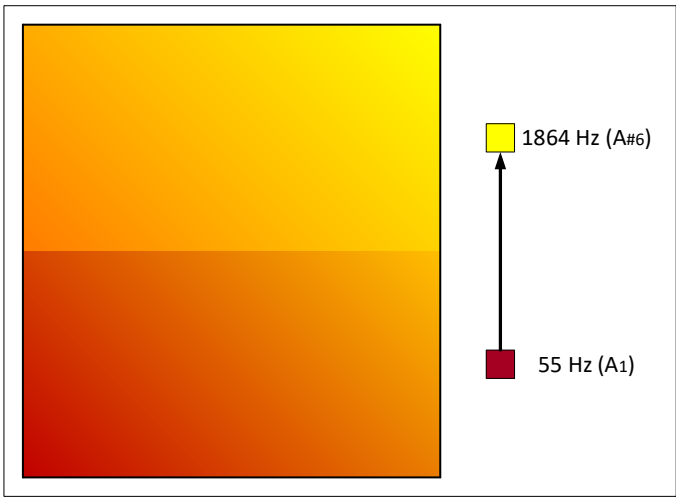

Abbildung 3: Der Frequenzverlauf der auditiven Repräsentation für ein Bild. Dabei ist zu sehen, wie die Frequenz sowohl vertikal als auch horizontal ansteigt. Rot stellt dabei die niedrigen Frequenzen und Gelb die hohen Frequenzen dar.

III. Realisierung

Startpunkt für das Programm ist das Menu (s. Abbildung 4). Dort kann ausgewählt werden, welche Quelle genutzt wird (s. Abbildung 2, Schritt 1). Da alle diese Quellen eine einheitliche Schnittstelle bedienen, beschreibe ich zuerst die einzelnen Quellen, um anschließend die weitere Berechnung und Ausgabe zu erläutern.
Als erste Auswählmöglichkeit ist die „Webcam“ gelistet. Die Webcam wird mithilfe der Java Bibliothek „Webcam Capture“ von Bartosz Firyn [7] eingebunden, da die Processing `Webcam` Bibliothek mit einigen Webcams nicht kompatibel ist. Nach Initialisierung der Webcam (s. `WebcamCapture` Z. 21ff) werden die Bilder von der Webcam abgefragt und in das `PImage`-Format von Processing umgewandelt, um für die weiteren Verarbeitungsschritte nutzbar zu sein (s. `WebcamCapture` Z. 30ff). Als weitere Auswahlmöglichkeit ist der „Choose“-Button dargestellt. Dieser öffnet einen Dialog, der dem Nutzer die Möglichkeit gibt, im Dateisystem nach einem beliebigen

Bild zu suchen. Dieses Bild wird anschließend ebenfalls als `PImage` abgespeichert (s. `SeeWithEars` Z. 231ff). Die letzte Auswahlmöglichkeit ist der „Fixed“-Button. Er bewirkt, dass aus einem Set von vier festen Bildern eines zufällig ausgesucht wird. Nach zwei Sekunden wird ein neues Bild ausgewählt. Der „Back“-Button führt den Nutzer zurück in den Menu-Bildschirm.
Nachdem die Bilder in das Programm geladen wurden, werden sie in Schritt 2 skaliert und mithilfe der Hilbert-Kurve neu sortiert (s. `Hilbert`). Die Skalierung des Bildes ist bedingt durch die Ordnung der Hilbert-Kurve. Je höher die Ordnung, desto größer ist auch die Anzahl an Pixeln die verarbeitet werden können. Als Standard ist eine Hilbert-Kurve erster Ordnung gewählt, sodass das eingehende Bild auf 2x2 Pixel runterskaliert wird, da eine Hilbert-Kurve erster Ordnung vier Ecken hat. Die höchste Auflösung ist 8x8 Pixel (Hilbert-Kurve 3. Ordnung). Höhere Auflösungen würden später zu einem starken Anstieg der Frequenzen führen, wodurch die Schwierigkeit für die Probanden enorm steigen würden, da die Differenzierbarkeit der Frequenzen entsprechend schwerer wird. Nachdem die Pixel sortiert wurden, werden die Quadranten I, II und IV vertauscht, um später den Frequenzverlauf aus Abbildung 3 zu entsprechen (s. `Hilbert` Z. 62ff). Während des Sortierungsprozesses werden zusätzlich alle derzeit überflüssigen Informationen des Bildes entfernt und ausschließlich die Helligkeit abgespeichert (s. `Hilbert` z.B. Z. 96). Diese Funktion sollte bei einer weiteren Überarbeitung des Programms in die `SoundProcessing`-Klasse verlagert werden, da sie thematisch besser in das Mapping von Bild zu Sound passt, als in die Sortierung. Dadurch würde die Konsistenz des Programms ansteigen.
Nachdem die Bilder akquiriert und sortiert sind, beginnt die Umwandlung der visuellen in die auditiven Informationen (s. Abbildung 2, Schritt 3). Dafür werden zuerst die Amplituden entsprechend der Helligkeit der Pixel berechnet (s. `Soundprocessing` Z. 45ff). Anschließend werden die zu nutzenden Frequenzen berechnet und den Pixeln zugeordnet (s. `Soundprocessing` Z. 65 und Z. 69).
In Schritt 4 werden dann die Sinus-Oszillatoren erstellt, die später den Ton erzeugen. Die Anzahl der Oszillatoren entspricht der Anzahl an Pixel im skalierten Bild, wird also indirekt durch die Ordnung der Hilbert-Kurve bestimmt. Jeder Oszillator bekommt eine Frequenz und die dazugehörige Amplitude zugewiesen (s. `Soundprocessing` Z. 67ff). Anschließend werden die Oszillatoren gestartet, sodass ein Ton entsteht, der die auditive Repräsentation des Bildes darstellt (Schritt 5).
Um den Benutzer auch ein visuelles Feedback zu geben, wird gleichzeitig das ursprüngliche Bild dargestellt, sodass eine klare Zuordnung zw. Bild und Ton zu ermöglichen (s. Abbildung 5)

IV. Fazit

Ziel des Projektes war die Umsetzung eines Systems zur auditiven Darstellung von Bildern. Aufbauend auf verschiedenen Studien wurde dafür ein System konzipiert, welches neue Ansätze in diesem Bereich verwirklicht. Vor allem die

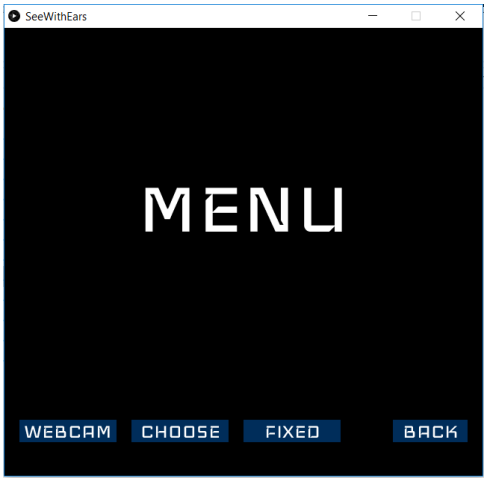

Abbildung 4: Das Menü des Programms. Im unteren Bereich sind vier Buttons zu sehen, mit denen das Programm gesteuert werden kann.

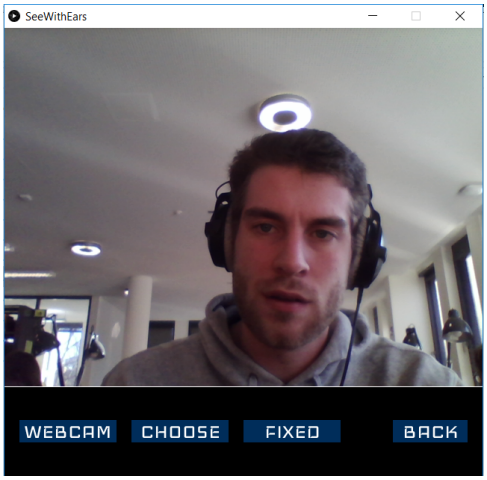

Abbildung 5: Darstellung von einem Bild in der Benutzeroberfläche. Hier: Webcambild. Gleichzeitig wird dem Nutzer die auditive Repräsentation des Bildes in Form eines Tons präsentiert.

Nutzung der Hilbert-Kurve zur Verbesserung der Trainingsmöglichkeiten auch für verschiedene Auflösungen ist ein wichtiger Bestandteil. Das System besteht aus fünf verschiedenen Teilbereichen. Der erste Bereich dient der Signalakquirierung, also der Erfassung eines Bildes, entweder von einer Webcam, von der Festplatte, oder aus einem Set von festgelegten Beispielbildern. Anschließend wird das Bild zur weiteren Verarbeitung skaliert und die Pixel mithilfe der Hilbert-Kurve sortiert. Dieser Ansatz unterscheidet das Projekt maßgeblich von den vorherigen Forschungen. In Schritt 3 werden die wichtigen Informationen aus dem Bild extrahiert, in diesem Fall also die Helligkeit der verschiedenen Pixel. Dies kann dann zusammen mit der Reihenfolge der neusortierten Pixel genutzt werden, um die Frequenzen und die Amplituden des Tons zu berechnen, der als auditive Repräsentation des Bildes agiert. Dieser wird dann zusammen mit dem genutzten Bild ausgegeben.
Da dieses System als Hilfsmittel für z.B. Blinde genutzt wer-

den bzw. die Forschung an eben diesen Hilfsmitteln unterstützen könnte, ist es unbedingt notwendig, Tests mit Probanden durchzuführen, um die Sinnhaftigkeit der neuen Ansätze zu überprüfen und quantitativ zu validieren. Dies ist bisher noch nicht geschehen. Bis zur Durchführung der Tests sollen jedoch noch einige Änderungen an dem Programm vorgenommen werden. Zum einen sollte die Auswahl der Tonlängen und damit der Framerate als Slider in die GUI eingefügt werden. Zusätzlich sollte auch die Größe der Hilbert Ordnung und damit die Anzahl an Frequenzen pro Ton in der Benutzeroberfläche angepasst werden können. Ein zusätzliche Erweiterung wäre die Nutzung von Soundsamples, anstatt vom Computer generierte Sinusschwingungen zu verwenden. Dadurch würde die Klangqualität erhöht und das Nutzererlebnis verbessert werden.

Im Rahmen des Projektes wurde deutlich, dass bei weiteren Softwareprojekten eine ausgedehntere Planungsphase mehr im Fokus liegen sollte. Wenn die Festlegung der nötigen Features und des strukturellen Aufbaus bereits vor Beginn der Umsetzung des Projektes abgeschlossen sind, verringert sich der Arbeitsaufwand bezüglich der Neustrukturierung des Programmcodes im Verlauf des Projektes enorm.

Dieses Programm kann als Ausgangspunkt für weitere Forschungen in dem Bereich der visuell-auditiven Substitution genutzt werden und ermöglicht darauf aufbauend die Realisierung von Folgeprojekten.

Anhang

Literatur

[1] Malika Auvray, Sylvain Hanneton und J Kevin O'Regan. „Learning to perceive with a visuo—auditory substitution system: localisation and object recognition with 'The Voice'". In: *Perception* 36.3 (2007), S. 416–430.

[2] Paul Bach-y-Rita u. a. „Vision substitution by tactile image projection". In: *Nature* 221.5184 (1969), S. 963–964.

[3] Ira H Bernstein und Barry A Edelstein. „Effects of some variations in auditory input upon visual choice reaction time." In: *Journal of experimental psychology* 87.2 (1971), S. 241.

[4] David Brown, Tom Macpherson und Jamie Ward. „Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device". In: *Perception* 40.9 (2011), S. 1120–1135.

[5] Christian Capelle u. a. „A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution". In: *IEEE Transactions on Biomedical Engineering* 45.10 (1998), S. 1279–1293.

[6] John Cronly-Dillon, Krishna Persaud und RPF Gregory. „The perception of visual images encoded in musical form: a study in cross-modality information transfer". In: *Proceedings of the Royal Society of London B: Biological Sciences* 266.1436 (1999), S. 2427–2433.

[7] Bartosz Firyn. *Webcam Capture*. http://webcam-capture.sarxos.pl/. 2017.

[8] David Hilbert. „Ueber die stetige Abbildung einer Line auf ein Flächenstück". In: *Mathematische Annalen* 38.3 (1891), S. 459–460.

[9] Charles Lenay u. a. „Sensory substitution: limits and perspectives". In: *Touching for knowing* (2003), S. 275–292.

[10] Lawrence E Marks. „On cross-modal similarity: Auditory–visual interactions in speeded discrimination." In: *Journal of Experimental Psychology: Human Perception and Performance* 13.3 (1987), S. 384.

[11] Peter BL Meijer. „An experimental system for auditory image representations". In: *IEEE transactions on biomedical engineering* 39.2 (1992), S. 112–121.

[12] Robert D Melara und Thomas P O'brien. „Interaction between synesthetically corresponding dimensions." In: *Journal of Experimental Psychology: General* 116.4 (1987), S. 323.

[13] Grant Sanderson. *The Hilbert Curve*. https://www.youtube.com/watch?v=3s7h2MHQtxc&t=641s. 2016.

Abbildungsverzeichnis