Künstliche Intelligenz in der Sprachverständlichkeit für Broadcast und Streaming

Davide Straninger[1]
(ds202@hdm-stuttgart.de)

[1]AM3, Hochschule der Medien, 70569 Stuttgart, Deutschland

Kurzfassung

Aufgrund von altersbedingtem Hörverlust haben ältere Menschen oft Schwierigkeiten, der Sprache im Fernsehen zu folgen. Doch nicht nur ältere Menschen leiden unter Hörverlust, denn dieser kann auch durch Genetik, Krankheit oder Unfälle bereits in jüngeren Jahren auftreten. Weiterhin ist die Sprache in manchen Medien und Streaming Plattformen auch mit gesundem Hörvermögen bereits schwierig zu verstehen und letztendlich können die Hörumgebung und das Abhörmedium ebenfalls einen negativen Einfluss haben. Deutsche Rundfunkanstalten erhalten jährlich viele Beschwerden wegen unzureichender Sprachverständlichkeit im Fernsehen und auf Mediatheken [2]. Die Technologie der KI-gestützten Signalseparation, insbesondere in Verbindung mit Next-Generation-Audio, verspricht eine mögliche Lösung dieser Problematik, auf welche im Folgenden näher eingegangen werden soll.

HOCHSCHULE DER MEDIEN

0 Problematik

Fernsehen, Streaming, Filme, Audio und Videomaterial sind ein Kulturgut und wichtige Informationsmedien unserer Zeit. Sie umgeben uns nahezu andauernd und machen einen immer weiter wachsenden Anteil unserer täglichen Mediennutzung aus [13]. Um an den gesellschaftlichen Entwicklungen teilzuhaben, ist es daher wichtig, die Sprache in diesen Medien für möglichst viele Menschen in der Gesellschaft verständlich zu machen.

Da sich die demographische Entwicklung in fast allen modernen Gesellschaften in Richtung eines höheren Durchschnittsalters verändert, und zunehmendes Alter oft mit Hörverlust einhergeht, wird dieses Problem in Zukunft noch wichtiger werden und muss von Forschung und Entwicklung adressiert werden [11]. Es besteht folglich die Notwendigkeit, die Sprachverständlichkeit in digitalen Medien zu verbessern.

Die Gründe für eine unzureichende Sprachverständlichkeit können vielfältig sein. So muss es nicht zwangsläufig an der hörenden Person liegen, sondern kann auch durch die Produktion und die Form der Distribution bedingt sein. Schauspieler können nuscheln, es kann schlecht aufgenommen worden sein, durch entweder fehlendes Können und Wissen oder durch den Aufnahmeort bedingt. Auch kann das Audiomaterial schlecht gemischt worden sein, durch entweder fehlendes Können und Wissen oder die kreative Entscheidung des Regisseurs. Weiterhin kann das Loudness-processing der Streaming Plattform negative Einflüsse haben und letztendlich können die Hörumgebung und das Abhörgerät auch einen Einfluss darauf haben.

1 Vorteile von KI in Verbindung mit Next-Generation-Audio

Zuerst zur Klärung des Begriffes „Next-Generation-Audio“ oder NGA. Dieser Begriff bezeichnet laut der European Broadcasting Union (EBU) keinen speziellen Audiocodec, sondern vielmehr eine Vielzahl von Technologien, mit deren Hilfe Audio im Broadcast-Bereich interaktiv, personalisierbar, skalierbar und immersiv produziert und übertragen werden kann [14]. Ein Beispiel hierfür wäre der am Fraunhofer IIS entwickelte MPEG-H Audiostandard, mit welchem sowohl 3D Mehrkanalaudio übertragen werden, als auch dem Endbenutzer die Möglichkeit gegeben werden kann, die Audiospur im Fernsehen z.B. hinsichtlich der Barrierefreiheit zu personalisieren [1].

Künstliche Intelligenz kann in Verbindung mit NGA zu einer Menge positiver Veränderungen des Produktions-, Distributions- und Konsumprozesses beitragen.

So kann durch die Verwendung künstlicher Intelligenz beispielsweise Zeit gespart werden, indem man Prozesse automatisieren und nach den eigenen Präferenzen durchführen lassen kann, für wel-

che Menschen im Durchschnitt mehr Arbeitszeit benötigen würden. Ein konkretes Beispiel hierfür wäre das sogenannte „Ducking“[1] bei Sprachmischungen für alternative Audiovisionen wie etwa der Audiodeskription[2].

Wo Zeit gespart werden kann, können auch Kosten gespart werden. Wenn ein Arbeitsprozessrozess einmal etabliert und automatisiert ist, kann er durch KI schneller erledigt werden, als ein Mensch das könnte.

Mit der KI-gestützten Signalseparation sind neben den Ersparnissen hinsichtlich Zeit und Kosten auch noch weitere Optionen offen. So ist es beispielsweise möglich, der/dem Endbenutzer*In die Möglichkeit zu geben, die Audiomischung nach den eigenen Präferenzen einzustellen. Es können dem Benutzer zu Hause Presets oder voreingestellte Wahlmöglichkeiten gegeben werden, die Mischung zwischen Sprache und Musik zu wählen oder auch bei entsprechend produzierten Live-Sendungen z.B. die eigene Mannschaft lauter zu hören, als die gegnerische [4].

Besonders für Hörgeschädigte Personen bietet sich eine barrierefreie Audioversion durch KI-gestützte Signalseparation an. Bereits fertig gemischtes Audiomaterial kann so hinsichtlich einer besseren Sprachverständlichkeit neu aufbereitet werden.

2 Ein prinzipieller Workflow

Damit KI-gestützte Signalseparation überhaupt eingesetzt werden kann, ist es sinnvoll, die verschiedenen Arten des Ausgangsmaterials zu differenzieren.

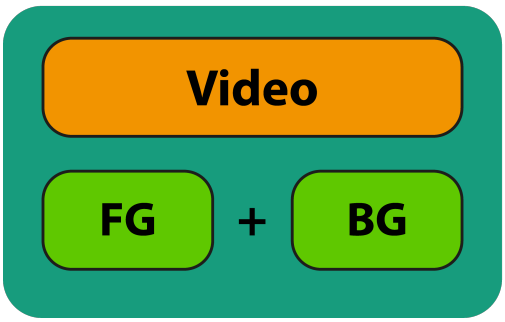

Abbildung 1: *Audiovisuelles Material mit einzeln verfügbaren Signalkomponenten*

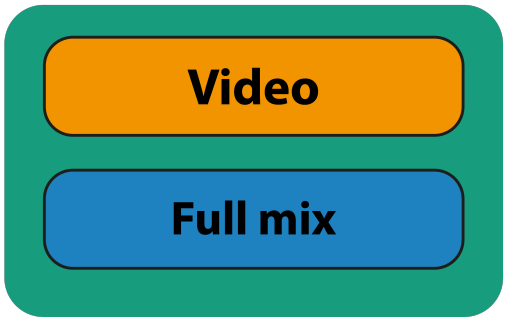

Abbildung 2: *Audiovisuelles Material mit dem Full-Mix ohne einzelne Signalkomponenten*

Unterschieden werden kann zwischen Ausgangsmaterial mit verfügbaren Signalkomponenten[3] oder Audiomaterial, bei

[1] Als „Ducking“ wird in der Audiofachsprache die zeitlich variable Lautstärke- oder Dynamikabsenkung eines Audiosignals verstanden, wenn ein weiteres Audiosignal als Trigger gleichzeitig aktiv ist.

2 Als Audiodeskription wird eine alternative Audiovision für visuell oder auditiv eingeschränkte Personen bezeichnet, bei der eine Stimme die Handlungen im Bild erläutert.

[3] In diesem Zusammenhang sind mit „Signalkomponenten“ die Sprache (auch Vordergrund/ Foreground oder FG) und alle weiteren nichtsprachlichen Signale wie Musik, Effekte und Geräusche (auch Hintergrund/ Background oder BG) gemeint.

welchem die Signalkomponenten nicht mehr einzeln vorliegen, sondern zu einem Full-Mix[4] zusammengefügt sind. Letztere Kategorie ist oft bei Archivmaterial im Fernsehen und natürlich auch auf Online-mediatheken und Streamingdiensten zu finden.

Wenn die Signalkomponenten bereits getrennt vorliegen, kann ein Neuronales Netz diese frei in der Lautstärke und Dynamik regeln und neu zusammenmischen. Die Absenkung des Hintergrundes kann dabei nach vorgegebenen Parametern wie etwa Attack, Hold oder Release erfolgen.

Wenn die Signalkomponenten allerdings nicht einzeln vorliegen, sondern zu einem Full-Mix zusammengemischt sind, ist es ohne die Hilfe von Algorithmen kaum möglich, etwa die Sprache aus dem Mix zu extrahieren und die Balance zwischen FG und BG neu zu definieren. An dieser Stelle kommt KI, beziehungsweise ein Neuronales Netz zum Einsatz. Dieses führt dann eine Signal-/ oder Dialogseparation durch und teilt den Full-Mix in Vordergrund und Hintergrund.

Mit den auf diese Weise separierten Komponenten kann dann eine sog. Dialog-Hervorhebung durchgeführt werden. Die meist zugrunde liegende Technik nennt sich Blindquellentrennung (engl. Blind-Source-Separation) [8]. Dabei schätzt der Algorithmus was Vordergrund und was Hintergrund ist und separiert diese Anteile dann in eigene Audiodateien. Diese können dann von einem weiteren Neuronalen Netz neu abgemischt und der Dialog hervorgehoben werden.

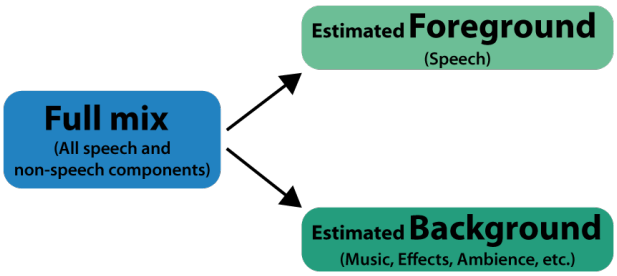

Abbildung 3: *Separation des Full-Mix in Foreground und Background*

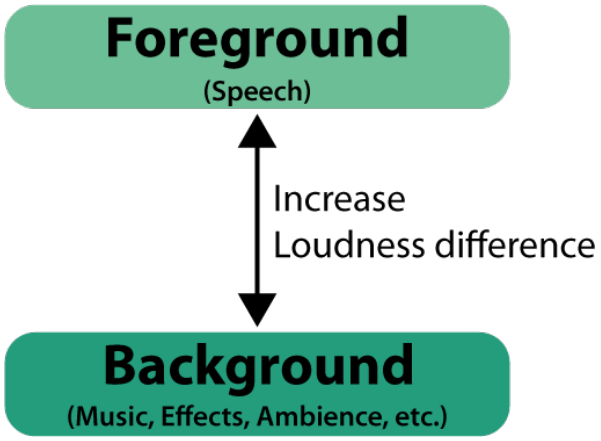

Abbildung 4: *Veränderung der Loudness Difference zwischen FG und BG zur Veränderung der Sprachverständlichkeit*

Hierbei wird dann meist nicht nur eine statische Absenkung des BG-Signals vorgenommen, sondern auch gleichzeitig ein zeitlich variables Ducking des BG-Signals durchgeführt.

Mit einem NGA-fähigen Übertragungscodec und entsprechender Technik in den Geräten der Endbenutzer kann man dann beispielsweise Presets zur Verfügung stel-

[4] Als Full-Mix wird ein Audiosignal bezeichnet, bei dem alle Signalkomponenten, wie etwa Sprache, Musik oder Effekte, zu einer Datei zusammengemischt sind.

len oder dem Consumer gar komplett freie Hand bei der Personalisierung des Audiomaterials lassen.

Trainiert wird das Neuronale Netzwerk für die Signalseparation in der Regel mit hochqualitativem Audiomaterial. Dafür benötigt es zunächst Audiospuren mit klarer, isolierter Sprache und Audiospuren mit dem Hintergrundsignal. Mit diesen Signalen wird das neuronale Netzwerk „gefüttert“. Anschließend werden die Einzelspuren zusammengemischt und das Netzwerk soll den Zustand herstellen, welcher bei den reinen Einzelspuren geherrscht hat. Das Resultat der Separation wird als eine Art Feedback mit der Information über den Grad des Erfolgs der Separation an das Netz zurückgegeben. Auf diese Weise kann es lernen und sich verbessern. Dieser Prozess muss mit vielen Stunden Audiomaterial vollzogen werden, um die Qualität des Ergebnisses zu verbessern.

3 Aktuelle Probleme

Die Separationsqualität, welche die Netzwerke aktuell erreichen können, ist bereits erstaunlich gut, jedoch noch immer weit davon entfernt, eine perfekte Quellentrennung zu erzielen. Meist kommt es zu starken Artefakten und auch zu sog. Ghost-Voicings[5] im separierten BG.

Wohingegen KI gute Ergebnisse bei statischen, nicht stark modulierenden BG-Signalen erzielt, ist spezielles Material, wie beispielsweise Stimme über Stimme oder Stimme über Musik, noch relativ kritisch. Auch stimmähnliche Geräusche wie z.B. Rabenvögel, sind schwieriger zu separieren, als andere Sounds.

Weiterhin könnte eine Tendenz zu erkennen sein, dass die besten Netzwerke nur von den größten Firmen mit guten finanziellen Mitteln erstellt werden, da es zeit- und kostenintensiv ist, an gutes Trainingsmaterial zu gelangen. Diese Problematik stellt sich jedoch nicht nur bei der KI-gestützten Signalseparation, sondern ist für den gesamten Bereich des Trainings von KI absehbar.

Eine weitere große Herausforderung stellt sich bei der Implementierung der Technologie in Rundfunkanstalten. Sowohl die Fernsehsender, als auch die Hersteller von entsprechenden Endgeräten müssen sich für eine aktive Verwendung von NGA-Systemen entscheiden. Dies geschieht allerdings nur, wenn die Akzeptanz der Zu-

[5] Als "Ghost Voicings“ werden Audioartefakte bezeichnet, die bei einer Blind-Source-Separation entstehen können und sich ähnlich wie Stimmen oder Dialog klingen.

schauer*Innen erlangt wird, da diese zusammen mit ökonomischen Gesichtspunkten die Hauptentscheidungsfaktoren darstellen, wenn es um die Implementierung der Techniken in den Produktionsablauf geht.

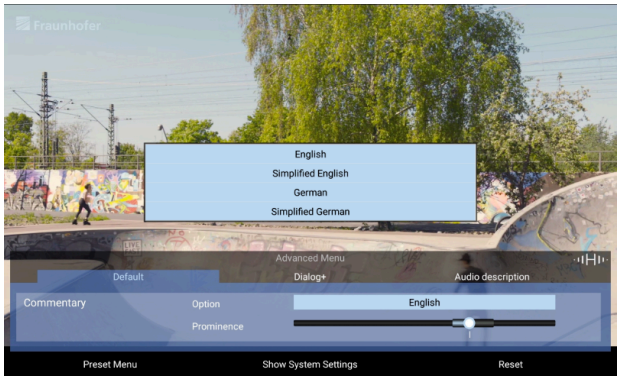

Abbildung 5: Grafische Benutzeroberfläche des MPEG-H Audiosystems zur Personalisierung der Sprachverständlichkeit

4 Bsp.1: Sprachverständlichkeit für Hörgeschädigte

Speziell in der Anwendung für Hörgeschädigte, seien es nun ältere Menschen mit einer Presbyakusis oder Menschen jeden Alters mit genetischem oder krankheitsbedingtem Hörverlust, findet KI-basierte Signaltrennung ihre Anwendung.

So soll beispielsweise MPEG-H[6] in der Lage sein, eine personalisierbare, barrierefreie Audioversion anzubieten, welche sich über ein grafisches Interface steuern lässt [1].

Die vom Fraunhofer IIS entwickelte Technologie zur Verbesserung der Sprachverständlichkeit nennt sich Dialog+ [8]. Es konnte in mehreren Untersuchungen mit kleineren Personengruppen gezeigt werden, dass die Möglichkeit, die Sprache zu personalisieren, bei den Zuschauer*Innen einen positiven Effekt bewirken kann und die gewünschten Erfolge erzielt.

So zeigt eine Arbeit beispielsweise einen positiven Effekt der zugrundeliegenden Technologie auf Personen über 65 Jahren [12]. Diese nutzten die Möglichkeit der Personalisierung und Vergrößerung der Loudness Difference intensiv und gaben auch an, durch die Möglichkeit der Personalisierung und Verbesserung der Sprachverständlichkeit eine größere Zufriedenheit zu verspüren, als ohne diese Möglichkeit.

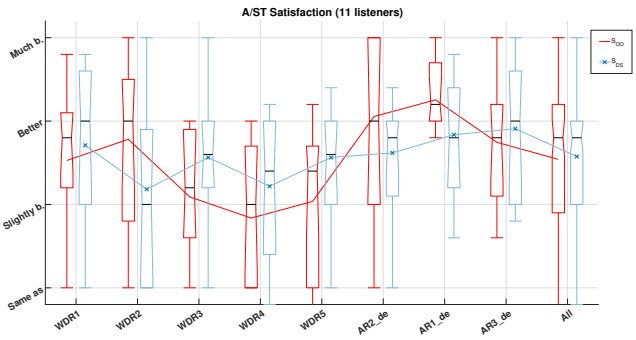

Abbildung 6: Durchschnittliche Zufriedenheitswerte der 11 Testpersonen bei unterschiedlichem Audiomaterial, welche alle im Bereich von „etwas besser“ bis „deutlich besser“ liegen

Tests zu dieser Technik wurden auch in größerer Form durchgeführt. So fand in Zusammenarbeit zwischen dem Fraunhofer IIS und dem Westdeutschen Rundfunk (WDR) z.B. eine Studie dazu im Internet statt, bei welcher die Testpersonen zwischen verschiedenen Audioversionen mit

[6] MPEG-H ist ein vom Fraunhofer IIS entwickeltes NGA-Audiosystem, welches unter Anderem 3D-Audio übertragen kann oder die Möglichkeit zur Personalisierung der Sprachverständlichkeit anbietet.

unterschiedlich stark abgesenktem Hintergrund anhören und bewerten konnten [10]. Letztendlich wird KI zur Verbesserung der Sprachverständlichkeit aber natürlich nicht nur von Fraunhofer erforscht und getestet, sondern findet auch bei anderen Herstellern und deren Systemen Anwendung. So z.B. auch im NEXTGEN TV oder Atsc 3.0. In letzterem dann unter dem Namen “Voice +” [5].

5 Bsp.2: Tools für die Postproduktion

Neben den Technologien, welche in NGA-Audiosystemen implementiert sind, gibt es auch kommerziell für den Privatgebrauch erwerbliche Software, welche erwähnt werden sollten, wenn es um KI-basierte Signalseparation geht.

Der iZotope RX8 Audioeditor ist ein in der Audio-Postproduktion gebräuchliches Tool, um Störgeräusche jeglicher Art entweder per Hand in einer Spektralansicht oder automatisiert per künstlicher Intelligenz zu entfernen [3]. Seit einiger Zeit sind auch Module zur Dialogverbesserung vorhanden, die eine Cloudbasierte, KI-gestützte Signalprozessierung verwenden, um die Sprache in Audiomaterial verständlicher zu machen und lästige Nebengeräusche zu entfernen.

Weitere Anwendung findet sich beispielsweise in der Software XTRAX Stems von Audionamix [9]. Hiermit kann man die einzelnen Stems[7] in einem Full Mix in der Laustärke regeln und somit auch die Sprache hervorheben.

Ein weiteres Beispiel von KI-basierter Dialogseparation, welches 2019 viel Aufmerksamkeit erregte, ist die Software “Spleeter” des Streamingdienstes “Deezer” [7]. Dieses Tool ist sogar open-source und steht auf GitHub zur freien Verfügung.

6 Fazit

KI-basierte Dialogseparation ist eine vielversprechende Technik, die in Zukunft in verschiedensten Szenarien eingesetzt werden könnte. Sei es im Fernsehen, auf Streamingplatformen oder z.B. auch im Flugzeug. Mit Hilfe von Presets, die automatisch das Endgerät oder die Umgebung erkennen oder frei einstellbarer Personalisierung kann man die Audiomischung nach den individuellen Präferenzen und dem eigenen Hörvermögen anpassen. Die Wirksamkeit dieser Technologie konnte bereits in Studien bewiesen werden und die Qualität der Algorithmen ist mitunter bereits sehr gut. Doch gibt es trotz allem noch aktuelle Probleme, welche durch weitere Forschung angegangen werden müssen. Nicht zuletzt kann KI-basierte Dialogverbesserung dabei helfen, das Thema der Inklusive voranzutreiben.

[7] Als „Stems“ werden in der Audiofachsprache die einzelnen Komponenten einer Audioproduktion, wie z.B. Audioeffekte, Geräusche, Stimme oder auch einzelne Instrumentengruppen, bezeichnet.

7 Quellen

7.1 Bildquellen

Abb.1: Eigene Darstellung

Abb.2: Eigene Darstellung

Abb.3: Eigene Darstellung

Abb.4: Eigene Darstellung

Abb.5: Simon, C., Torcoli, M. & Paulus, J. (2019). MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming. arXiv.org, 6. http://arxiv.org/abs/1909.11549

Abb.6: Straninger, D. (2020). Dialogue Enhancement in Object-based Audio - Evaluating the Benefit on People above 65. arXiv.org, 48. https://arxiv.org/abs/2006.14282

7.1 Literaturquellen

[1] Fernsehen in Deutschland [Internet]. Statista. [zitiert 6. August 2021]. Verfügbar unter: https://de.statista.com/statistik/studie/id/6383/dokument/fernsehen-in-deutschland-statista-dossier/

[2] Xtrax stems [Internet]. Audionamix. [zitiert 6. August 2021]. Verfügbar unter: https://audionamix.com/xtrax-stems/

[3] NextGen TV [Internet]. NEXTGEN TV is the Future of TV. Consumer Technology Association; [zitiert 6. August 2021]. Verfügbar unter: https://www.cta.tech/Membership/Member-Groups/Video-Division/NEXTGEN-TV

[4] Spleeter [Internet]. [zitiert 6. August 2021]. Verfügbar unter: https://research.deezer.com/projects/spleeter.html

[5] Eberhardt M, Eberhard M, Hergenröder E, Kiefer M, Matejka D, Siegfried A, u. a. Sprachverständlichkeit im Fernsehen Empfehlungen für Programm und Technik. ARD and ZDF; 2014 Juli S. 15. (Empfehlungen für Programm und Technik). Report No.: 3.

[6] EBU. WHY BROADCASTERS NEED AN OPEN, CODEC-INDEPENDENT WORKFLOW FOR NGA PRODUCTION DEPLOYMENT. Geneva: EBU; 2019 Jan S. 13. (Operating Eurovision and Euroradio). Report No.: 45.

[7] Fraunhofer iis und wdr testen dialog+ – fraunhofer audio blog [Blog Post]. [zitiert 6. August 2021]. Verfügbar unter: https://www.audioblog.iis.fraunhofer.com/de/fraunhoferiis-wdr-dialogplus

[8] Grewe Y, Simon C, Scuda U. Producing Next Generation Audio using the MPEG-H TV Audio System. BEITC. 2018;7.

[9] RX 8 Overview | The industry standard for audio repair [Internet]. iZotope. [zitiert 6. August 2021]. Verfügbar unter: https://www.izotope.com/en/products/rx.html

[10] WDR. TV-Dialoge besser verstehen: WDR testet mit seinem Publikum neues Verfahren [Internet]. 2020 [zitiert 6. August 2021]. Verfügbar unter: https://www1.wdr.de/unternehmen/der-wdr/unternehmen/dialog-plus-sprachverstaenlichkeit-fernsehen-102.html

[11] Simon C, Torcoli M, Paulus J. Mpeg-h audio for improving accessibility in broadcasting and streaming. arXiv:190911549 [cs, eess] [Internet]. 25. September 2019 [zitiert 6. August 2021]; Verfügbar unter: http://arxiv.org/abs/1909.11549

[12] Statistisches Bundesamt Wiesbaden. Bevölkerung im Wandel - Annahmen und Ergebnisse der 14. koordinierten Bevölkerungsvorausberechnung. Wiesbaden: Statistisches Bundesamt; 2019 Juni S. 71. (Bevölkerung im Wandel). Report No.: 14.

[13] Straninger D. Dialogue Enhancement in Object-based Audio - Evaluating the Benefit on People above 65. arXiv:200614282 [cs, eess] [Internet]. 25. Juni 2020 [zitiert 6. August 2021]; Verfügbar unter: http://arxiv.org/abs/2006.14282

[14] Wohlmayr M. Blind Source Separation : Eine Einführung [Internet]. Graz; [zitiert 6. August 2021]. Verfügbar unter: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.1280