# Audio-Browsing mit elastischen Schiebereglern im E-Learning

Wolfgang Hürst

Institut für Informatik
Albert-Ludwigs-Universität Freiburg
huerst@informatik.uni-freiburg.de

**Abstract:** Dieser Beitrag gibt zunächst eine Übersicht über verschiedene Verfahren zum Audio-Browsing, d.h. zur schnellen und einfachen Navigation in Audiodaten, und diskutiert ihre Relevanz für den Einsatz in der Lehre, zum Beispiel dem Selbststudium mit Vorlesungsaufzeichnungen. Dabei werden zwei Situationen als besonders nützlich und sinnvoll identifiziert: Das kontinuierliche Abspielen mit einer fixen, beschleunigten Geschwindigkeit sowie die kurzzeitige Beschleunigung der Abspielgeschwindigkeit. Darauf aufbauend wird im zweiten Teil ein Interaktionskonzept zur Navigation in Audiodaten vorgestellt, das auf dem Prinzip der so genannten „Elastic Interfaces" beruht. Die vorgeschlagene Realisierung unterstützt eben diese beiden Arten des Audio-Browsings auf besonders einfache und intuitive Art und Weise, weshalb sie sich insbesondere für den Einsatz in e-Learning Anwendungen eignet.

## 1 Einführung

Digitalisierte Tondokumente finden in vermehrtem Maße Verwendung beim Lehren und Lernen. Ein besonderes Beispiel mit einer großen und kontinuierlich zunehmenden Verbreitung sind Vorlesungsaufzeichnungen, die meist mit entsprechenden Verfahren automatisch erstellt und den Studenten beispielsweise zum Selbststudium oder zur Nachbereitung zur Verfügung gestellt werden [LMT04]. Derartige Aufzeichnungen umfassen in der Regel die Stimme des Vortragenden (oft als *Audio-Strom* bezeichnet), die präsentierten Unterrichtsmaterialien (Folien, etc.) inklusive der darauf gemachten Annotationen (auch *Whiteboard-Strom* genannt) sowie ggfs. eine Videoaufzeichnung des Dozenten (*Video-Strom*). Der Audio-Strom wird dabei im Allgemeinen als die wichtigste Informationsquelle angesehen [GB97, LMT04] Ferner belegen Studien im Zusammenhang mit der Verwendung derartiger Dokumente im regulären Studienalltag (siehe z.B. [ZH02]) eine sehr selektive Nutzung: Aufgezeichnete Vorlesungen werden nicht linear als ganzes betrachtet, sondern nur in ausgewählten Teilen wiedergegeben. Vergleichbar dem Arbeiten und Lernen mit einem Lehrbuch werden uninteressant erscheinende oder bereits anderweitig behandelte Passagen nur kurz überflogen oder komplett übersprungen, wichtige und schwer verständliche Stellen werden wiederholt gehört usw. Aus dieser Beobachtung ergeben sich bestimmte Anforderungen an die Abspiel-Software, mit der die betreffenden Dokumente betrachtet werden. Studenten

müssen beispielsweise in der Lage sein, den Inhalt schnell zu überfliegen, wenig relevante Stellen zu überspringen, einfach auf bestimmte Positionen zurückzusetzen, um komplizierte Inhalte wiederholt anzuhören usw.

Im Multimedia-Bereich werden für eine derartige Navigation und Interaktion oft auch die englischen Begriffe *Browsing* und *Skimming* verwendet. Dieses Browsing geschieht bei zeitsynchronen, kontinuierlichen multimedialen Datenströmen meist anhand der visuellen Information: Verkleinerte und zeitlich sortierte Darstellungen von Folien (*Thumbnails*) werden als Navigationshilfen benutzt, anhand des visuellen Feedbacks beim Bewegen eines Sliders entlang der Zeitachse werden bestimmte Inhalte klassifiziert, irrelevante Stellen schnell übersprungen und besonders interessante Positionen direkt angesteuert, etc. Gerade im Zusammenhang mit aufgezeichneten Vorträgen reicht eine derartige „visuelle Navigation" jedoch oft nicht aus, da sich der Hauptinformationsgehalt bei Vorlesungsaufzeichnungen, wie bereits erwähnt, meist im Audiostrom befindet. Zwar enthält eine Folie oft eine gute Zusammenfassung des aktuellen Inhalts des Audiosignals, jedoch werden Nebenbemerkungen und Querverweise, Abweichungen vom eigentlichen Thema, usw. nicht erfasst, obwohl gerade diese Dinge für eine Navigation oder Suche in den Daten häufig besonders wichtig sind. Ferner ist eine lineare Übereinstimmung des zeitlichen Ablaufs des Audiosignals mit der Anordnung des Textes auf einer Folie nicht immer gegeben. Insbesondere bei lange aufliegenden Folien, die keine animierten Objekte oder während des Vortrags hinzugefügte Annotationen enthalten, lassen sich daher häufig aus der rein visuellen Information nicht immer Rückschlüsse auf den aktuellen Inhalt des Audiosignals ziehen, auch wenn zwischen beiden ein klarer Zusammenhang besteht. Neben Verfahren zur visuellen Navigation sind somit auch Ansätze zum Browsing in Audiodaten gerade im Zusammenhang mit Lehrmaterial besonders wichtig.

Dieser Beitrag verfolgt daher zwei wesentliche Ziele: Zum einen wird eine Übersicht über verschiedene Ansätze zum Audio-Browsing präsentiert. Aufgrund entsprechender Studien wird ihre Relevanz und potentielle Bedeutung für das e-Learning diskutiert, und die für dieses Szenario sinnvollen Ansätze werden identifiziert. (*Kapitel 2*). Motiviert durch die derart spezifizierten Anforderungen wird ein konkretes Beispiel eines erweiterten Schnittstellenentwurfs eingeführt (*Kapitel 3*). Die damit realisierten Interaktionsmöglichkeiten zielen insbesondere darauf ab, eben diese Anforderungen auf einfache und intuitive Art und Weise zu unterstützen, weshalb die vorgeschlagene Realisierung gerade für e-Learning-Anwendungen besonders geeignet erscheint.

## 2 Ansätze für Audio-Browsing und Bezug zur Lehre

**Visualisierung des Audiosignals**. Wie bereits im vorangehenden Kapitel erwähnt, werden häufig die zeitsynchronen visuellen Datenströme für eine Navigation in Multimedia-Dokumenten herangezogen, da Audiosignale über die Zeit hinweg „vergänglich" sind. Visuelle Datenströme sind zwar auch linear, bestehen jedoch aus statischen Grundeinheiten (einzelne Bilder oder *Frames*) und eignen sich daher wesentlich besser für eine interaktive Navigation in den Daten. Ein beliebter Ansatz zum Browsing von Audio- und insbesondere Sprachdaten besteht daher in einer

Visualisierung des Audiosignals. Bei allgemeinen Audiodaten wird hierbei meist Metainformation visualisiert, wie zum Beispiel der Typ (Musik, Stille, Sprache, etc.) oder eine genauere Spezifikation (Sprecher-ID, Musiktitel und Interpret, etc.). Bei den bei Vortragsaufzeichnungen generell vorliegenden Sprachsignalen eines einzigen Sprechers wird häufig versucht, mittels automatischer Spracherkennung eine textuelle Beschreibung des Sprachsignals, ein sog. *Transcript*, zu generieren und entsprechend zu visualisieren [Hü04]). Eine völlig fehlerfreie Transformation des kontinuierlichen Sprachsignals in eine statische Repräsentation ist jedoch in der Regel nicht möglich. Ferner geht durch den damit verbundenen Medienwechsel im Sprachsignal enthaltene Information verloren, da gedruckter Text nicht die Ausdrucksstärke gesprochener Sprache besitzt. Merkmale wie Betonung, Sprechgeschwindigkeit usw. lassen sich damit (wenn überhaupt) nur schwer darstellen [Ar94]. Deren Wichtigkeit für das Browsing wurde jedoch gerade im Zusammenhang mit Vortragsaufzeichnungen [He00] nachgewiesen. Wird der Text mittels automatischer Spracherkennung erstellt, fehlen ferner auch die für ein Überfliegen des Inhalts hilfreichen Satzzeichen sowie weitere Strukturinformation wie Absätze, Einrückungen, Fettdruck wichtiger Worte usw.

Für den Einsatz in e-Learning-Anwendungen sind derartige Verfahren daher wohl wenig geeignet. Interessanter erscheinen hier vielmehr Ansätze, bei denen eine Navigation durch direktes Anhören des gegebenenfalls modifizierten Audiosignals erfolgt. Diese beruhen in der Regel auf zwei Konzepten: dem zeitkomprimierten sowie dem inhaltskomprimierten Abspielen des Audiosignals. Bei der zeitkomprimierten Wiedergabe erfolgt ein Browsing durch das schnellere Abspielen des Sprachsignals. Bei der inhaltskomprimierten Wiedergabe werden einzelne, als weniger wichtig klassifizierte Teile des Sprachsignals, zum Beispiel Sprechpausen, beim Abspielen übersprungen. Aus diesem Grund spricht man häufig auch von einer nichtlinearen zeitkomprimierten Wiedergabe. Auf beide Ansätze wird im Folgenden genauer eingegangen. Die Betrachtung beschränkt sich dabei auf Sprachsignale.

**Zeitkomprimierte Wiedergabe**. Durch den Einsatz signalverarbeitender Algorithmen lässt sich erreichen, dass ein Sprachsignal auch bei einer höheren oder niedrigeren Abspielgeschwindigkeit noch „natürlich“, der echten Stimme des jeweiligen Sprechers entsprechend klingt. Derartige Verfahren werden meist unter den Begriffen *Time Stretching* oder *Time Scaling* zusammengefasst, welche jedoch neben der insbesondere für das schnelle Überfliegen des Inhalts vor allem interessanten Zeitkomprimierung (*Time Compression*) auch ein verlangsamtes Abspielen (*Time Expansion*) umfassen. Erfolgt die Wiedergabe innerhalb gewisser Geschwindigkeitsgrenzen, so lässt sie sich gewinnbringend für das schnelle Auffinden einer bestimmten Information einsetzen, ähnlich wie beim Überfliegen des Inhalts eines visuellen Datenstroms durch ein beschleunigtes Abspielen. Für eine detaillierte Beschreibung von Algorithmen zur zeitkomprimierten Wiedergabe unter Beibehaltung der Charakteristika des Sprachsignals sei auf [Hü05] verwiesen.

Arons [Ar94, Ar97] unterscheidet im Zusammenhang mit der schnelleren Wiedergabe von Sprachsignalen zwischen „Verständlichkeit“ (*Intelligibility*) als der Fähigkeit, jedes einzelne Wort verstehen zu können, und „Verständnis“ (*Comprehension*) als der Fähigkeit, den entsprechenden Inhalt noch verstehen und aufnehmen zu können. Bezug

nehmend auf ältere Arbeiten wird berichtet, dass bei zusammenhängender, kontinuierlicher Sprache ein Verständnis bei einer bis zum 2,0-fachen schnelleren Abspielgeschwindigkeit möglich ist, während bei einer höheren Abspielgeschwindigkeit kritische, nicht-redundante Information verloren geht. He und Gupta [HG01] verweisen ebenfalls auf ältere Arbeiten, die von einer „komfortablen" Abspielgeschwindigkeit im Bereich der 1,25- bis 1,4-fachen Normalgeschwindigkeit berichten, präsentieren jedoch Studien aus denen ein wesentlich höherer Wert von einer 1,6- bis 1,7-fachen Beschleunigung resultiert. Basierend auf einer Vergleichsstudie mit unterschiedlichen Algorithmen zur zeitkomprimierten Wiedergabe schlussfolgern die Autoren auch, dass dieser Wert möglicherweise die obere Grenze der menschlichen Aufnahmefähigkeit darstellt. Neben diesen allgemeinen Untersuchungen, sind für das e-Learning natürlich vor allem Studien interessant, die im Zusammenhang mit aufgezeichneten Video-Vorlesungen durchgeführt wurden. Galbraith und Spencer [GS01] berichten beispielsweise von einer durchschnittlich 1,6-fachen Abspielgeschwindigkeit, wobei die meisten Teilnehmer der 256 Studenten umfassenden Studie nach eigenen Angaben die 1,3- bis 1,8-fache Wiedergabegeschwindigkeit bevorzugten. Harrigan [Ha00] sowie Li et al. [Li00] berichten, ebenfalls im Zusammenhang mit Video-Vorlesungen, von bevorzugten Abspielraten vom 1,4- bzw. 1,23-fachen der normalen Abspielgeschwindigkeit. Amir et al. [Am00] präsentieren eine Studie, bei der explizit das Verständnis in Abhängigkeit von der Abspielgeschwindigkeit getestet wurde. Während bei Radionachrichtensendungen ein maximales Verständnis bei der 1,4-fachen Normalgeschwindigkeit erreicht wurde, lag die für das Verständnis optimale Wiedergaberate bei einem aufgezeichneten Vortrag bei einer 0,84-fachen Geschwindigkeit, war also langsamer als die normale Abspielgeschwindigkeit. Dieser Wert resultierte sowohl aus den subjektiven Benutzereinschätzungen als auch aus einer objektiven Bewertung durch einen neutralen Beobachter. Auch wenn sich dieses Ergebnis aufgrund des geringen Umfangs der Studie nicht verallgemeinern lässt und einer weiteren Überprüfung bedarf, zeigt es doch, dass auch ein verlangsamtes Abspielen sinnvoll für den Lernprozess sein kann, insbesondere bei neuem oder besonders schwierigem Material sowie bei fremdsprachigen Vorträgen (ein nicht unwesentlicher Anteil der Studienteilnehmer besaßen eine andere Muttersprache als der jeweils Vortragende). Stifelman et al. [SAS01] sowie Galbraith und Spencer [GS01] berichten dagegen von Aussagen der Teilnehmer ihrer jeweiligen Studien, gemäß derer ein beschleunigtes Abspielen eine höhere Konzentration erfordert, was den Lernerfolg in Einzelfällen gemäß der subjektiven Aussagen verbessert hat. Arons [Ar94] verweist auf eine Studie von Sticht [St69], bei der sich das zweimalige Hören von Lehrmaterial mit doppelter Abspielgeschwindigkeit als effizienter herausgestellt hat als das einmalige Hören mit Normalgeschwindigkeit.

Als erste Beobachtung lässt sich somit festhalten, dass das beschleunigte (oder ggfs. verlangsamte) Abspielen aufgezeichneter Vorträge zwar sinnvoll erscheint, eine einheitliche, „optimale" Geschwindigkeit hierfür jedoch nicht existiert. Systeme, bei denen die zeitkomprimierte Wiedergabe im Zusammenhang mit Lehrmaterial eingesetzt wird, sind zum Beispiel DUKES [Ha99] und SPECIAL [Ha00] sowie das sogenannte Audio Notebook [SAS01] Neben universitären Systemen und Forschungsprototypen halten Verfahren zur zeitkomprimierten Wiedergabe vermehrt auch Einzug in kommerzielle Programme, wie zum Beispiel den Windows Media Player [MI04].

Die im Vorangehenden beschriebenen Studien beschäftigten sich vor allem mit der beschleunigten kontinuierlichen Wiedergabe einzelner Sprachdokumente, den dafür sinnvollen Geschwindigkeitsgrenzen und den möglichen Auswirkungen auf den Lernerfolg. Die selektive Nutzung aufgezeichneter Vorträge, die sich aus den in Kapitel 1 beschriebenen Studien ergeben hat, macht jedoch deutlich, dass auch eine kurzzeitige Navigation in den Daten, z.B. mit dem Ziel, eine als irrelevant erscheinende Stelle schneller überfliegen zu können, im e-Learning-Kontext von besonderer Bedeutung ist. Verschiedene Studien bestätigen, dass das zeitkomprimierte Abspielen gerade auch für ein derartiges schnelles „Überfliegen" bzw. Lokalisieren bestimmter Information interessante Perspektiven bietet. Bei einer derartigen „Suche durch Browsing" ist es nämlich häufig nicht zwingend notwendig, den kompletten Inhalt zu verstehen. Stattdessen kommt es meist „nur" auf eine Klassifikation und Identifikation relevanter Teile und Information an. Aus diesem Grund sind hierfür gegebenenfalls auch deutlich höhere Geschwindigkeiten denkbar. Unsere eigenen informellen Studien mit aufgezeichneten Radionachrichtensendungen ergaben in diesem Zusammenhang z.B. bei einem routinierten Benutzer Geschwindigkeitsobergrenzen vom bis zu 2,5-fachen der Normalgeschwindigkeit. Ferner berichtet [Ar97] von Trainingseffekten beim zeitkomprimierten Abspielen von Audiodaten und demzufolge einem verbesserten Verständnis auch bei relativ hoher Abspielgeschwindigkeit. Die dort zitierte Aussage von Beasley und Maki [BM76], dass sich Benutzer nach einem mindestens 30-minütigen Anhören eines zeitkomprimierten Audiostroms beim Rückfall auf eine normale Abspielgeschwindigkeit „unwohl" fühlten, konnte durch informellen Tests von uns bestätigt werden.

Auch wenn vergleichbare Studien im Zusammenhang mit e-Learning-Anwendungen nach unserem Kenntnisstand leider nicht existieren, scheint eine Übertragung dieser Ergebnisse doch nahe liegend. Als zweite Beobachtung lässt sich daher festhalten, dass im e-Learning-Kontext auch eine kurzzeitige Beschleunigung (oder ggfs. Verlangsamung) der Abspielgeschwindigkeit sinnvoll erscheint und deshalb von der Schnittstelle unterstützt werden sollte.

**Automatische Segmentierung zur inhaltskomprimierten Wiedergabe**. Neben einer reinen Beschleunigung der Wiedergabe lässt sich ein schnelleres Abspielen eines Sprachsignals natürlich auch erreichen, indem die darin enthaltenen Pausen automatisch identifiziert und entfernt werden. Allerdings enthalten Pausen oft syntaktische und semantische Hinweise, so dass es nicht immer sinnvoll erscheint, sie komplett zu entfernen [Ar94]. Gerade diese implizit enthaltene Information lässt sich jedoch nutzen, um automatisch Strukturinformation über den jeweiligen Dokumentinhalt zu bekommen, die dann für eine inhaltskomprimierte Wiedergabe genutzt werden kann. Im Zusammenhang mit der Navigation in Audiodaten ist man daher häufig nicht nur an einer Segmentierung der Daten interessiert, sondern auch daran, „wichtigere" Stellen, die den betreffenden Inhalt besonders gut repräsentieren, automatisch zu bestimmen. Durch das automatische Überspringen weniger relevanter Teile des Signals kann sowohl das Abspielen als auch die Suche nach wichtiger Information beschleunigt werden.

Hinweise auf besonders relevante Teile oder Einschnitte im Inhalt liefern oft die im Signal enthaltenen Pausen sowie die Betonung des jeweiligen Sprechers. Längere

Sprechpausen geben häufig einen Hinweis auf den Beginn eines neuen Satzes, Gedankens oder Unterthemas und ihre Dauer korreliert oft mit ihrem Typ und ihrer Wichtigkeit [Ar97]. Atempausen beim Vorlesen von Text dauern in der Regel um die 400 ms. Bei Spontansprache sind Pausen generell häufig länger und die Wortrate, das heißt die Anzahl gesprochener Worte pro Minute, ist niedriger als beim Vorlesen von Text [Ar94]. Arons [Ar94] unterscheidet allgemein zwischen relativ kurzen „Verzögerungen“ (*Hesitation Pauses*), die für gewöhnlich eine Dauer zwischen 200 und 250 ms haben, und bewusst gemachten „Verbindungspausen“ (*Juncture Pauses*), die wesentlich länger sind und im Schnitt 500 bis 1000 ms dauern. Gerade letztere sind für das Verständnis sehr wichtig, lassen aber auch Rückschlüsse auf die inhaltliche Struktur des Signals zu. Pfeiffer [Pf01] nutzt die Pausenlänge beispielsweise zur Generierung einer Segmentierung auf „Phrasen-Ebene“ sowie auf „Szenen-Ebene“, wobei Phrasen in der Regel einzelne Sätze umfassen, während Szenen größere semantische Einheiten repräsentieren. Dabei liefert eine Pausengröße zwischen 100 und 400 ms die besten Ergebnisse bei einer Segmentierung in einzelne Phrasen, während für eine szenen-basierte Unterteilung des Sprachsignals Pausengrößen zwischen 400 ms und 2.150 ms verwendet werden.

Neben der Pausendauer ist bei Sprachsignalen auch eine Veränderung der Betonung ein Indiz für den Beginn eines neuen Themas oder dafür, dass der Sprecher der betreffenden Stelle einen besonderen Nachdruck verleihen will. Laut Arons [Ar94] geht beispielsweise die Einführung eines neuen (Unter-)Themas oft mit einer Erhöhung der Stimmlage einher, während am Ende eines Satzes die Tonhöhe häufig reduziert wird. Ähnliche Aussagen werden von Hirschberg und Grosz gemacht [HG92, GH92]. Durch eine automatische Analyse der Betonung lässt sich somit nicht nur eine Segmentierung des Sprachsignals generieren, sondern auch Information über die Relevanz einzelner Segmente ableiten.

Während bei der zeitkomprimierten Wiedergabe diverse Studien mit aufgezeichneten Vorlesungen und Vorträgen existieren (vgl. vorangehenden Abschnitt), wurden in der Literatur bei der Einführung und Überprüfung der pausen- und betonungs-basierten Verfahren zur automatischen Segmentierung häufig andere Daten verwendet, wie zum Beispiel Nachrichtensendungen. Um zu verifizieren, ob bzw. in welchem Maße sich die dortigen Aussagen über die Interpretation der Pausenlänge bzw. einer Veränderung der Betonung auf die in unserem Szenario vorkommenden Sprachsignale, das heißt die aufgezeichnete Stimme eines Vortragenden, übertragen lassen, wurden daher von uns verschiedene, detaillierte Untersuchungen durchgeführt [Di00]. Leider bestätigte sich hier der Verdacht, dass sich die gemachten Aussagen in ihrer Allgemeingültigkeit nicht auf Vorlesungsaufzeichnungen übertragen lassen. Sowohl eine pausen- als auch eine betonungs-basierte Segmentierung führen bei Vorlesungsaufzeichnungen in der Regel nicht zu den gewünschten Ergebnissen, d.h. nicht zu einer Unterteilung des Sprachsignals in sinnvolle inhaltliche Einheiten. Während Sprachdaten wie zum Beispiel Nachrichtensendungen in der Regel „wohlformuliert“ sind, d. h. von geschulten Sprechern mit sinnvollen Pausen und Betonungen gesprochen werden, gleichen die Audiomitschnitte von Vorlesungen eher einer Spontansprache: Pausen sind nicht immer inhaltlich motiviert, sondern ergeben sich auch aus kurzfristigen Überlegungen heraus,

Betonungen werden nicht zwangsläufig zu Beginn eines neuen Absatzes gemacht, „Äh“'s und „Hmm“'s unterbrechen den Redefluss usw.

Als Fazit lässt sich somit festhalten, dass von den verschiedenen Ansätzen – der textuellen Repräsentation des Sprachsignals, der beschleunigten Wiedergabe mittels Time Scaling sowie der automatischen Segmentierung zur Navigationsunterstützung und Komprimierung des Signals – nur die Modifikation der Abspielgeschwindigkeit Erfolg versprechend erscheint. Des Weiteren kann man schlussfolgern, dass sowohl eine andauernde Veränderung der Abspielgeschwindigkeit auf einen höheren oder niedrigeren Wert sinnvoll sein kann, als auch eine kurzfristige Manipulation, wie zum Beispiel eine temporäre Beschleunigung, um eine weniger relevante Stelle schnell zu überspringen.

## 3 Audio-Browsing mit Elastic Interfaces

Im Folgenden wird ein von uns entwickelter Schnittstellenentwurf vorgestellt, der den zuvor identifizierten Anforderungen für das Audio-Browsing im Zusammenhang mit e-Learning Anwendungen besser gerecht wird. Leider unterstützen heutige Medien-Player derartige Interaktionsmechanismen oft nur unzureichend, da sich ihr Schnittstellenentwurf noch vornehmlich an der traditionellen „Tape-Recorder“ Metapher orientiert. Dies gilt insbesondere für die Möglichkeit einer kurzzeitigen Modifikation der Abspielgeschwindigkeit. Einige Systeme erlauben zwar deren stufenloses Verstellen, zum Beispiel mit einer Art Schieberegler, wie in Abbildung 1 dargestellt. Dieser ist jedoch vornehmlich für die dauerhafte Einstellung einer bestimmten Geschwindigkeit geeignet. Ein ständiges Verändern der Abspielgeschwindigkeit, wie es bei einer interaktiven Navigation zur Suche in einer Audiodatei mitunter erforderlich ist, lässt sich damit nur schwer durchführen. Auch eine nur kurzzeitige Beschleunigung, um beispielsweise eine uninteressante Stelle schnell zu überspringen, erfordert immer das wiederholte Eingreifen eines Benutzers, wenn die betreffenden Bereiche überschritten sind und wieder zur vorherigen Abspielgeschwindigkeit zurückgekehrt werden soll.

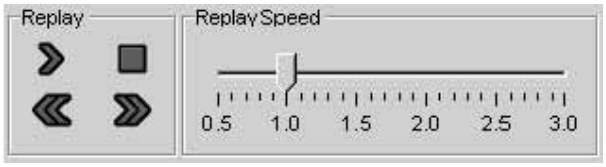


Abbildung 1: Schieberegler zur Geschwindigkeitsmanipulation.

Besser für eine flexible interaktive Navigation in den Daten geeignet wäre in einer solchen Situation wahrscheinlich ein mit der Dokumentlänge assoziierter, zeitbasierter Schieberegler oder *Slider*, der durch Bewegen des Slider-Elements eine Navigation entlang der Zeitachse ermöglicht. Derartige Schnittstellen haben sich zur schnellen und flexiblen interaktiven Navigation in visuellen Datenströmen, insbesondere auch im Zusammenhang mit Lehrmaterial, bewährt [HG04]: Liefert die entsprechende Abspiel-Software während der Bewegung des Sliders ein visuelles Feedback in Echtzeit, ermöglicht dies dem Benutzer eine schnelle Klassifikation des Inhalts, es erlaubt es auf

eine flexible und einfache Art, zwischen verschiedenen Bewegungsgeschwindigkeiten im Dokument zu wechseln, schnell auf bestimmte Stellen zurückzusetzen usw. Leider ist ein derartiges, „unmittelbares“ Feedback im Zusammenhang mit akustischen Datenströmen nicht möglich. Während sich visuelle Medienströme in der Regel aus statischen Grundeinheiten zusammensetzen – einzelnen (Stand-)Bildern oder *Frames* – bestehen digitale Tondokumente aus *Samples*, die für sich genommen keine sinntragende Einheit bilden, sondern nur in einer hinreichend langen Folge eine wahrnehm- und interpretierbare Information vermitteln. Durch das Abspielen einer zusammengehörigen Folge aufeinander folgender Samples beim Bewegen des Sliders würde jedoch die Synchronität zwischen Slider-Element und Wiedergabe verletzt. Gängige Medien-Player, die ein akustisches Feedback während der Navigation entlang der Zeitachse liefern, verzichten daher in der Regel auf eben diese Synchronität – eine Vorgehensweise, die zwar hilfreich ist, um einen groben Überblick über den Inhalt zu bekommen, für eine detaillierte Suche und flexible Navigation auf unterschiedlichen Detailebenen jedoch kritisch erscheint. Im Folgenden wird daher ein Ansatz vorgestellt, der eine flexible Navigation im Dokument bei zeitsynchroner Audio-Wiedergabe ermöglicht, indem die Slider-Bewegungen, die ein Benutzer durchführen kann, derart eingeschränkt werden, dass ein zeitsynchrones und für die Suche und Navigation hilfreiches Audio-Feedback möglich ist.

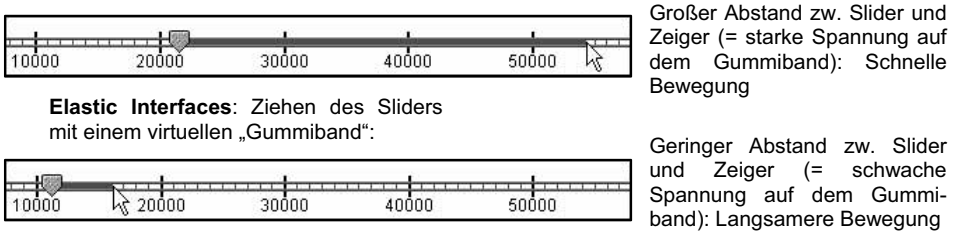


Abbildung 2: Illustration des Konzepts der Elastic Interfaces.

**Elastic Interfaces**. Die hier umgesetzte Idee zur Realisierung einer flexiblen Navigation durch Audio-Dateien mit zeitsynchronem, akustischem Feedback beruht auf dem Grundkonzept der so genannten *Elastic Interfaces*. Dieses wurde 1995 im Zusammenhang mit der Navigation in statischen Daten, wie z.B. Texten und Bildern, eingeführt [MKB95]. Bei einem Elastic Interface wird das betreffende Slider-Element zur Navigation nicht direkt bewegt, sondern entlang einer Verbindungslinie – einem „elastischen Gummiband“ – zwischen Mauszeiger und Slider gezogen (vgl. Abb. 2). Die Geschwindigkeit, mit der der Slider den Mauszeigerbewegungen folgt, ist direkt abhängig von der Entfernung zwischen diesen beiden Objekten: Wird dieser Abstand größer, nimmt die Spannung auf dem „Gummiband“ zu, der Slider bewegt sich daher schneller (bzw. die Geschwindigkeit, mit der man sich durch das Dokument bewegt, nimmt zu). Verringert sich der Abstand, nimmt die Spannung ab, und die Slider- bzw. Scrolling-Geschwindigkeit reduziert sich entsprechend. Ein Benutzer kann sich somit sehr schnell und sehr langsam durch ein Dokument bewegen, indem er den Mauszeiger vom Slider entfernt bzw. ihm entsprechend annähert. Intern wird der Abstand zwischen den beiden Objekten (Mauszeiger und Slider) mittels einer Mapping-Funktion auf eine Scrolling-Geschwindigkeit abgebildet, wie in Abb. 3a illustriert. Motivation für diese

Art Schnittstelle in den entsprechenden Originalarbeiten von Masui et al. [MKB95, Ma98] war es, den Benutzern auch eine Navigation auf einer beliebigen Detailebene in großen Dokumenten zu ermöglichen; eine Funktionalität, die von herkömmlichen Slidern nicht geliefert werden kann, da sie eine Navigation im Sub-Pixel-Bereich erfordern würde. Durch das Konzept der Elastic Interfaces wird dieses Problem behoben und eine flexible Navigation auf unterschiedlichen Granularitätsstufen ermöglicht.

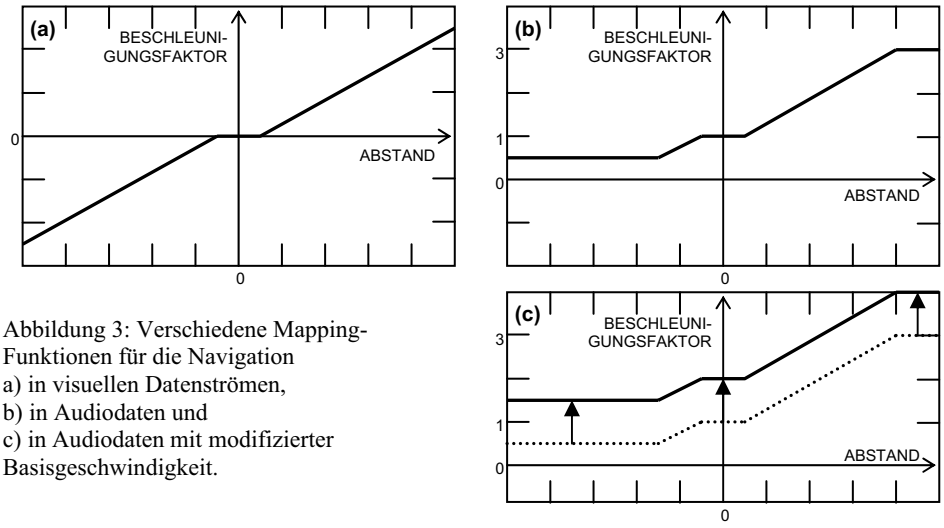


Abbildung 3: Verschiedene Mapping-Funktionen für die Navigation
a) in visuellen Datenströmen,
b) in Audiodaten und
c) in Audiodaten mit modifizierter Basisgeschwindigkeit.

**Der Elastic Audio Slider**. Durch die in Abbildung 3a illustrierte Mapping-Funktion zwischen Zeiger-Slider-Abstand und Scroll- bzw. Navigationsgeschwindigkeit wird, wie erwähnt, eine feingranulare Bewegung in den betreffenden Dokumenten ermöglicht, die mit normalen Slidern nicht realisierbar wäre. Mit diesem Ansatz wird folglich ein sehr großes Geschwindigkeitsspektrum abgedeckt – von sehr langsamer Navigation auf feingranularer Ebene bis hin zur sehr schnellen Bewegung durch ein Dokument. Der Wechsel zwischen unterschiedlichen Geschwindigkeiten erfolgt jedoch nicht abrupt wie bei der Bewegung eines normalen Sliders, sondern kontinuierlich. Der von uns vorgeschlagene Elastic Audio Slider nutzt diese Eigenschaft aus, um mit Hilfe einer geeignet definierten Mapping-Funktion die Bewegung des Sliders derart einzuschränken, dass eine zeitsynchrone Audio-Wiedergabe möglich ist.

Im vorangehenden Kapitel wurden bereits die Grenzen für Abspielgeschwindigkeiten in unterschiedlichen Szenarien besprochen. Es konnten zwar keine einheitlichen Werte identifiziert werden, die Studien mit aufgezeichneten Vorträgen sowie unsere eigenen diesbezüglichen Erfahrungen legen jedoch nahe, dass bei der 1,4- bis 1,8-fachen Geschwindigkeit ist oft noch ein komplettes Verständnis möglich ist. Bei 2,0- bis 2,5-facher Beschleunigung lässt sich der Inhalt noch gut klassifizieren. Je nach Dokument und Routine des Benutzers ist dies sogar noch bei einer höheren Geschwindigkeit möglich. Unsere eigenen Tests lieferten eine obere Schranke von 3,0 als absolutes Maximum für einen sinnvollen Einsatz mit Vorlesungsaufzeichnungen. Ein verlangsamtes Abspielen kann in gewissen Situationen ebenfalls sinnvoll sein. Durch

eine entsprechende Anpassung der Mapping-Funktion eines Elastic Slider, wie sie in Abbildung 3b illustriert ist, lässt sich damit ein *Elastic Audio Slider* realisieren der vom Grundprinzip her ähnlich funktioniert, wie sein visuelles Gegenstück: Zieht man den Slider nach rechts, erhöht sich die Abspielgeschwindigkeit entsprechend bis zu einer sinnvollen Obergrenze des maximal 3-fachen der normalen Wiedergabe. Eine Bewegung nach links verringert die Abspielgeschwindigkeit entsprechend. Abbildung 4 veranschaulicht die Visualisierung, die in unserer Implementierung umgesetzt wurde.

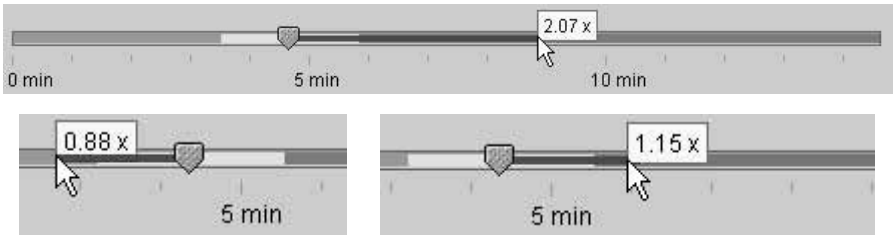


Abbildung 4: Implementierung des Elastic Audio Sliders (verschiedene Beispiele mit einer Bescheunigung bzw. Verlangsamung auf das 2,07-fache, das 0,88-fache bzw. das 1,15-fache der normalen Abspielgeschwindigkeit).

Die interaktive Manipulation der Abspielgeschwindigkeit eignet sich vor allem zur kurzfristigen Beschleunigung oder Verlangsamung der Wiedergabe, also für Situationen, wie sie gerade beim Lernen häufiger vorkommen. Da es durchaus auch sinnvoll sein kann, die Abspielgeschwindigkeit konstant auf einen anderen Wert zu setzen (z.B. eine Beschleunigung um 1,2 bei relativ langsamen Sprechern oder eine Verlangsamung auf 0,8 bei fremdsprachlichen Vorträgen), ist der Elastic Audio Slider mit einem normalen Regler zur Einstellung der Abspielgeschwindigkeit (vgl. Abb. 1) gekoppelt. Wird mit diesem eine neue Abspielgeschwindigkeit – im folgenden *Basisgeschwindigkeit* genannt – eingestellt, wird auch die Mapping-Funktion entsprechend angepasst (siehe Abb. 3c, die das Beispiel einer Anpassung der Basisgeschwindigkeit auf die doppelte Abspielgeschwindigkeit illustriert). In verschiedenen Laborstudien wurden unsere Thesen bezüglich Flexibilität und Intuitivität des Elastic Audio Sliders und seine Eignung für die interaktive Navigation in Audiodaten bestätigt. Insbesondere die Möglichkeit, kurzfristig zu beschleunigen und durch ein einfaches Loslassen sofort zur zuvor eingestellten Basisgeschwindigkeit zurück zu kehren ohne diese erneut einstellen zu müssen, wurde von verschiedenen Teilnehmern als besonderer Vorteil angesehen. Die Studien wurden meist mit Nachrichtensendungen durchgeführt. Eine umfassende Studie zum Einsatz des Elastic Audio Sliders im realen Lehrbetrieb wird zurzeit (Sommersemester 2006) durchgeführt. Hierfür wurde in einer Kooperation mit der imc AG der an unserer Fakultät intensiv eingesetzte Player der LECTURNITY® Software [im06] um verschiedene Möglichkeiten des Audio-Browsings sowie einen automatischen Logging-Mechanismus erweitert. In einer ca. zweimonatigen Studie wird nun untersucht, ob und wenn ja wie die entsprechenden Mechanismen von den Studenten tatsächlich genutzt werden. Aufgrund der in Kapitel 2 präsentierten Diskussion über die Nützlichkeit unterschiedlicher Audio-Browsing Verfahren vermuten wir, dass sich die These einer besonderen Eignung des Elastic Audio Sliders für e-Learning Anwendungen bestätigen wird.

## 4 Zusammenfassung

Ausgehend von der Aussage, dass Verfahren zum Audio-Browsing gerade im Zusammenhang mit digitalem Lernmaterial besonders wichtig sind, wurden im Vorangehenden zunächst verschiedene Ansätze hierfür diskutiert und bezüglich ihrer Relevanz für das Lehr- und Lernszenario untersucht. Neben einer konstanten Veränderung der Abspielgeschwindigkeit über einen längeren Zeitraum hinweg (zum Beispiel einer Verlangsamung, wenn der Vortrag in einer Fremdsprache gehalten wurde, oder einer Beschleunigung, wenn eine Vorlesung zu Wiederholungszwecken abgespielt wird), haben sich dabei auch kurzfristige Erhöhungen und Erniedrigungen der Abspielgeschwindigkeit als nützlich erwiesen (zum Beispiel wenn ein wenig relevanter Teil der Daten schnell überflogen werden soll oder eine vergleichsweise schwierige Stelle ein besonders sorgfältiges Durcharbeiten erfordert). Während Ersteres, d.h. die konstante Wiedergabe mit einer höheren oder niedrigeren Abspielgeschwindigkeit, von den Benutzerschnittstellen gängiger Medien-Player zum Beispiel durch Schieberegler zum Einstellen des Beschleunigungsfaktors oft gut unterstützt wird, lässt sich eine flexible, kurzfristige Manipulation damit in der Regel nur schwer durchführen. Deshalb wurde von uns im zweiten Teil dieses Beitrags das Elastic Audio Slider Interface vorgestellt – ein Schnittstellenkonzept, das auf dem Prinzip der Elastic Interfaces beruht und eben diese Funktionalität bietet. Die damit umgesetzte Art der Interaktion fügt sich nahtlos in den Entwurf bestehender Schnittstellen ein. Diese werden daher nicht ersetzt, sondern um eine neue Funktionalität erweitert. Dem Benutzer wird somit ein größeres Spektrum zur Navigation in Audiodaten geboten, das insbesondere die beiden zuvor identifizierten, für das e-Learning typischen Interaktivitäten umfasst.

## Literaturverzeichnis


[Am00] Amir, A. et al.: Using audio time scale modification for video browsing. Video Use in Office and Education, 33rd Hawaii Int. Conf. On System Sciences (HICCS 2000), Januar 2000.

[Ar94] Arons, B.: Interactively skimming recorded speech. PhD thesis, Massachusetts Institute of Technology, 1994.

[Ar94] Arons, B.: Speechskimmer: A system for interactively skimming recorded speech. ACM Transactions on Computer-Human Interaction, 4(1):3-38, 1997.

[BM76] Beasley, D. S.; Maki, J. E.: Time- and Frequencey-Altered Speech. Chapter 12 in (Lass, N. J., Hrsg.): “Contemporary Issues in Experimental Phonetics.”, New York: Academic Press, S. 419-458, 1976.

[Di00] Dick, J.: Analyse und Indizierung von Audio-Dateien für das Information Retrieval in Multimedia-Dokumenten. Diplomarbeit, Albert-Ludwigs-Universität Freiburg, April 2000.

[GS01] Galbraith, J.; Spencer, S.: Variable speech playback (VSP) and asynchronous video-based instruction: Evaluating leaner feedback. Industry whitepaper. Erhältlich unter http://www.enounce.com/docs/BYUPaper020319.pdf, 2001.

[GB97] Gemmell, J.; Bell, G.: Noncollaborative telepresentations come of age. Communications of the ACM, Vol. 40, No. 4, April 1997.

[GH92] Grosz, B.; Hirschberg, J. Some intonational characteristics of discourse structure. Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP 1992), Vol. 1, S. 429-432, Personal Publishing Ltd., Oktober 1992.

[Ha99] Harrigan, K.: DUKES: Creating real-time variable-speed speech for use in educational multimedia. Proceedings of ED-MEDIA – World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 1999), S. 1360-1361, AACE, Juni 1999.

[Ha00] Harrigan, K.: The SPECIAL system: Searching time-compressed digital video lectures. JRCE, Fall 2000, 33(1), 2000.

[He00] He, L. et al.: Comparing presentation summaries: Slides vs. reading vs. listening. Proceedings of the Conference on Human Factors in Computing Systems (ACM CHI 2000), S. 177-184, ACM Press, April 2000.

[HG01] He, L.; Gupta, A.: Exploring benefits of non-linear time compression. Proceedings of the 9th ACM International Conference on Multimedia (ACM MM 2001), S. 382-391, ACM Press, September/Oktober 2001.

[HG92] Hirschberg, J.; Grosz, B.: Intonational features of local and global discourse structure. Proceedings of the Speech and Natural Language Workshop, S. 441-446, Defence Advanced Research Projects Agency (DARPA), Morgan Kaufmann Publishers, Februar 1992.

[Hü04] Hürst, W.: User Interfaces for Speech-Based Retrieval of Lecture Recordings. Proceedings of ED-MEDIA – World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2004), AACE, Lugano, Switzerland, June 2004.

[HG04] Hürst, W.; Götz, G.: Interface Issues for Interactive Navigation and Browsing of Recorded Lectures and Presentations. Proceedings of ED-MEDIA – World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA 2004), AACE, Lugano, Switzerland, June 2004.

{Hü05] Hürst, W.: Multimediale Informationssuche in Vortrags- und Vorlesungsaufzeichnungen. Promotionsschrift, Albert-Ludwigs-Universität Freiburg, Februar 2005.

[imc06] imc AG: LECTURNITY® http://www.lecturnity.de/, 2006

[LMT04] Lauer, T.; Müller, R.; Trahasch, S.: Learning with lecture recordings: key issues for end-users. Proceedings of ICALT 2004, Joensuu, Finnland.

[Li00] Li, F. C. et al.: Browsing digital video. Proceedings of the Conference on Human Factors in Computing Systems (ACM CHI 2000), S. 169-176, ACM Press, April 2000.

[MKB95] Masui, T., Kashiwagi, K., Borden, G.R. IV: Elastic graphical interfaces for precise data manipulation. ACM CHI 1995 (conference companion), S. 143-144, ACM Press.

[Ma98] Masui, T.: LensBar – Visualization for browsing and filtering large lists of data. Proceedings of the 1998 IEEE Symposium on Information Visualization (INFOVIS 1998), pp. 113-120, IEEE Computer Society, 1998.

[Mi04] Microsoft: What's new in Windows Media Player 9 series, http://www.microsoft.com/windowsxp/windowsmediaplayer/9series/whatsnew.asp, 2004

[Pf01] Pfeiffer, S.: Pause concepts for audio segmentation at different semantic levels. Proceedings of the 9th ACM international conference on Multimedia (ACM MM 2001), S. 187-193, ACM Press, September/Oktober 2001.

[St69] Sticht, T. G.: Comprehension of repeated time-compressed recordings. The Journal of Experimental Education, 37(4):60-62, 1969.

[SAS01] Stifelman, L.; Arons B.; Schmandt, C.: The Audio Notebook – Paper and pen interaction with structured speech. Proceedings of the Conference on Human Factors in Computing Systems (ACM CHI 2001), S. 182-189, ACM Press, März/April 2001.

[ZH02] Zupancic, B.; Horz, H.: Lecture recordings and its use in a traditional university course. Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education (ITICSE 2002).