ASDF: Ein XML Format zur Beschreibung von virtuellen 3D-Audioszenen

Matthias Geier, Jens Ahrens und Sascha Spors

Deutsche Telekom Laboratories, TU Berlin, 10587 Berlin, Deutschland, E-Mail: Matthias.Geier@telekom.de

Einleitung

Räumliche Schallwiedergabe wird mittlerweile für viele Anwendungen genutzt und ist auf vielen Systemen verfügbar. Diese unterscheiden sich durch die Art der Wiedergabe, durch die Größe, durch den Verwendungszweck und viele andere Faktoren. Deshalb werden meistens auch für die Speicherung von räumlichen Audio-Inhalten völlig unterschiedliche Dateiformate verwendet. Dieser Umstand lässt es leider nicht zu, dass Audio-Szenen ohne erheblichen Mehraufwand auf unterschiedlichen Wiedergabesystemen abgespielt werden können.

Mit einem gemeinsamen Dateiformat könnte man diese problematische Situation verbessern und man würde den Austausch von virtuellen Audio-Szenen wie zum Beispiel räumlichen Kompositionen und Demos zwischen verschiedenen Wiedergabeorten und -systemen verbessern. Außerdem wäre eine Evaluierung von verschiedenen Systemen mit den gleichen Szenen-Daten möglich.

In diesem Artikel werden die Anforderungen an ein solches Format vorgestellt, es wird erörtert, warum einige bestehende Formate nicht geeignet erscheinen und es wird ein Entwurf für ein Dateiformat gezeigt, das die gegebenen Anforderungen erfüllen soll: das *Audio Scene Description Format (ASDF)*.

Anforderungen

Sowohl einfache *statische Szenen* als auch *dynamische Szenen* mit beliebigen Quellenbewegungen sollen möglich sein. Außerdem soll der Benutzer auf kontrollierte Weise mit der Szene *interagieren* können.

Quellen sollen zwar beliebig im dreidimensionalen Raum positioniert werden können, da aber viele Beschallungssysteme existieren, die auf die horizontale Ebene beschränkt sind, soll die Angabe von zweidimensionalen Positionen vereinfacht möglich sein.

Das Format soll *einfach zu verwenden* aber auch *einfach zu implementieren* sein, denn es wird sich nur als Austauschformat durchsetzen können, wenn es von vielen Nutzern implementiert wird. Es soll ohne spezielle Editor-Software *lesbar, veränderbar und generierbar* sein.

Das Format soll erweiterbar sein, damit auch zukünftige Entwicklungen in den Wiedergabemethoden berücksichtigt werden können.

Das ASDF

Das hier vorgestellte Format wurde entwickelt um die genannten Anforderungen zu erfüllen. Virtuelle Szenen werden in Text-Dateien gespeichert (mit Verweisen auf Audiodateien in etablierten binären Formaten), daher kann man sie mit einem gewöhnlichen Text-Editor bearbeiten.

Durch die Verwendung von XML (eXtensible Markup Language) [1] ist das Format sowohl von Menschen lesbar als auch flexibel erweiterbar. In der Praxis wird das Format bereits in der Wiedergabe-Software *SoundScape Renderer* [2] eingesetzt und es ist zu hoffen, dass es in Zukunft auch auf weiteren Systemen implementiert wird.

Quellen-Parameter

Virtuelle Quellen haben verschiedene Parameter, die in der Szenen-Beschreibung angegeben werden und die auch deterministisch oder interaktiv verändert werden können. Es müssen nicht alle denkbaren Parameter bei der Konzeption des Formates berücksichtigt werden, da das Format beliebig erweiterbar ist.

Relevante Parameter sind neben Position und Richtung der Schallquelle auch der Typ (zB Punktquelle, ebene Welle) und die Abstrahlcharakteristik. Einzelne Eigenschaften können fixiert werden, um die möglichen Benutzerinteraktionen einzuschränken.

Audio-Daten

Die Audio-Daten, die von den virtuellen Quellen abgespielt werden, können vielerlei Ursprungs sein. Der einfachste Fall ist eine lokal abgespeicherte mono Audio-Datei. Es können aber auch mehrkanalige Dateien angegeben werden. Die direkte Verwendung eines Einganges der Soundkarte ist natürlich auch möglich. Zusätzlich können auch Netzwerk-Streams angegeben werden.

Die Audio-Daten können unabhängig von der Quellenbewegung gestartet und gestoppt werden, des Weiteren können Audiodateien in einer „playlist“ hintereinander abgespielt werden.

Beispiel

Auf der nächsten Seite folgt ein Beispiel für eine Szene mit bewegten Quellen. Das Koordinatensystem ist mit `"RFT"` (right, front, top) festgelegt. Das bedeutet, dass die positive x-Achse nach rechts ausgerichtet ist, die y-Achse nach vorne und die z-Achse nach oben. Es handelt sich also um ein rechtshändiges Koordinatensystem.

In der Beispiel-Szene befinden sich zwei Schallquellen, denen jeweils sowohl eine Audio-Datei zugewiesen wird, als auch ein Quellen-Typ (Punktquelle) und eine Ausgangsposition im Raum. Dieses Beispiel ist auf die horizontale Ebene beschränkt, deswegen wird die dritte Koordinate einfach weggelassen. Bei der zweiten Audio-Datei wird

noch angegeben, dass sie nicht automatisch zum Beginn der Szene abgespielt wird; sie wird erst später gestartet.

Der `<score>`-Teil beinhaltet den dynamischen Teil der Szene. Zuerst wird die erste Quelle entlang einer Trajektorie bewegt, die Bewegung wird 8,4-mal wiederholt. 33 Sekunden nach dem Beginn wird dann die zweite Audio-Datei gestartet. Es ist kein explizites Ende angegeben, deswegen endet die Szene, sobald beide Audio-Dateien fertig abgespielt wurden.

```
<?xml version="1.0" encoding="utf-8"?>
<asdf version="0.2" coordinates="RFT">
  <scene volume="3" name="Moving Sources">
    <source id="source_1" model="point" position="0.4 1.5">
      <audiofile src="audio/one.wav"/>
    </source>
    <source id="source_2" model="point" position="-2 2">
      <audiofile id="file_two" src="audio/two.wav" start="manual"/>
    </source>
  </scene>
  <score>
    <par>
      <animate target="source_1">
        <trajectory id="tr01" repeat="8.4">
          <n p="5 5" t="5s"/> <n p="10 0" t="9s"/>
          <n p="5 -5" t="12s"/> <n p="0 0" t="18s"/>
        </trajectory>
      </animate>
      <seq>
        <wait dur="33s"/>
        <start target="file_two"/>
      </seq>
    </par>
  </score>
</asdf>
```

Alternative Formate

Es existieren bereits Dateiformate die die eingangs genannten Anforderungen zumindest teilweise erfüllen. In den folgenden Absätzen werden drei Formate kurz beschrieben und dabei die Aspekte hervorgehoben, in denen sie von den Anforderungen abweichen.

X3D/VRML

VRML (Virtual Reality Modeling Language) ist eine Beschreibungssprache für 3D-Grafik, die vor allem im Internet weit verbreitet ist und seit der Version 2.0 von der ISO standardisiert ist. Ihr Nachfolger, ebenfalls ISO-Standard, der neben der ursprünglichen Syntax auch eine XML Syntax und ein binäres Format bereitstellt, ist *X3D (eXtensible 3D)*. Beide Formate sind primär für 3D-Grafik-Modellierung gedacht, ermöglichen aber auch die Platzierung von Sounds im Raum [3].

Der Aufbau der beiden Formate basiert auf einem Szenen-Graphen-Modell. Dies ist eine Baum-Struktur in der alle Objekte hierarchisch angeordnet sind. Die tatsächliche Position und Orientierung eines Objektes berechnet sich unter Berücksichtigung aller im Szenen-Graphen übergeordneten geometrischen Transformationen. Dies ist eine sehr praktische und bewährte Methode um 3D-Modelle, die aus einer enormen Menge an simplen geometrischen Objekten aufgebaut sind, abzuspeichern. Da allerdings klingende Objekte meist nur durch eine einzige oder einige wenige Quellen modelliert werden, ist der Aufwand eines Szenen-Graphen für reine Audio-Szenen nicht gerechtfertigt. Geometrische Transformationen können einfach auf einzelne Objekte oder auf (nicht hierarchische) Gruppen von Objekten angewendet werden.

MPEG-4 AudioBIFS

Der ISO-Standard MPEG-4 beinhaltet unter dem Namen *BInary Format for Scenes (BIFS)* die komplette Funktionalität von VRML und ermöglicht zusätzlich das Streaming aller Audio- und Szenen-Daten. Der Umfang der Audio-Funktionen wird mit *AudioBIFS* stark erweitert [4]. Unter Anderem können auch akustisch wirksame Flächen und perzeptive Raum-Parameter angegeben werden.

MPEG-4 hat einen enormen Funktionsumfang, der für reine Audio-Szenen zum Großteil nicht verwendet würde. Aufgrund der Komplexität des Standards ist er auch sehr aufwendig zu implementieren. Da VRML zur Gänze in BIFS enthalten ist, gelten auch die oben genannten Einschränkungen bezüglich Szenen-Graphen.

SMIL

SMIL (Synchronized Multimedia Integration Language) [5] ist ein XML-basiertes Dateiformat zur Steuerung und Synchronisation von Audio-, Video- und 2D-Grafik-Inhalten und wird wie das englische Wort „smile“ ausgesprochen. Das Format spezifiziert nicht die eigentlichen Audio- und Video-Daten sondern organisiert nur die zeitliche Abfolge und die Positionierung auf dem Bildschirm. Dreidimensionale Positionierung und Bewegungspfade sind nicht möglich, deswegen eignet sich das Format auch nicht für dreidimensionale Audio-Szenen.

Ausblick

Das *Audio Scene Description Format* befindet sich noch in einem frühen Entwicklungsstadium, es ist allerdings wichtig, dass bereits jetzt die Anforderungen von potentiellen Nutzern berücksichtigt werden. Außerdem müssen noch viele Details (wie zB Raum-Eigenschaften, Skalierung von Szenen) gemeinsam diskutiert werden. Interessierte werden gebeten, sich an die oben genannte E-Mail-Adresse zu wenden.

Literatur

[1] World Wide Web Consortium. *eXtensible Markup Language (XML 1.0, Fourth Edition)*, Aug. 2006. `http://www.w3.org/TR/xml/`.

[2] M. Geier, J. Ahrens und S. Spors. The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *124th AES Convention*. Amsterdam, Niederlande, Mai 2008.

[3] Web3D Consortium. *eXtensible 3D (X3D)*, 2004. `http://www.web3d.org/x3d/`.

[4] R. Väänänen und J. Huopaniemi. Advanced AudioBIFS: Virtual acoustics modeling in MPEG-4 scene description. *IEEE Trans. Multimedia*, 6(5):661–675, Okt. 2004.

[5] World Wide Web Consortium. *Synchronized Multimedia Integration Language (SMIL 2.1)*, Dez. 2005. `http://www.w3.org/TR/SMIL2/`.