COSMOS의 3D 콘텐츠 음향정보 자동등록 기술

지수미, 권순일, 백성욱*
세종대학교 디지털콘텐츠학과
e-mail:sbaik@sejong.ac.kr

Audio Information Authoring Technology for 3D Contents of COSMOS

Su Mi Ji, Soonil Kwon, Sung Wook Baik
Dept of Digital Contents, Sejong University

요 약

COSMOS (COntentS Making Omnipotent System)는 컴퓨터 게임이나 3차원 애니메이션 제작이 가능하도록 그래픽 랜더링, 특수효과, 물리엔진, 인공지능 엔진 등의 기능을 갖춘 범용성 3차원 콘텐츠 저작 시스템이며, 무엇보다도 직관적인 인터페이스 기능을 통해 사용자의 편리성을 제공해 준다. 본 논문은 COSMOS에서 음향 정보를 자동으로 3D 콘텐츠 구성 요소에 배합될 수 있도록 하는 기술에 대한 내용이다. 본 기술의 도입을 통해 COSMOS에서는 사용자의 의성어 소리를 인식하여, 그 의미에 적합한 디지털 사운드를 검색한 후에 사용자의 의도에 맞추어 변환하여 이와 관련된 콘텐츠 구성 요소와 일치시켜줌으로써 보다 직관적으로 콘텐츠 저작 기능을 제공할 수 있다.

1. 서론

COSMOS(COntentS Making Omnipotent System) [1]는 인터랙티브 콘텐츠를 위한 지능형 3D 콘텐츠 저작 시스템이며, 본 시스템을 통해 그래픽 렌더링, 인공지능, 물리효과, 특수효과를 포함한 G3 Engine의 기능을 기반으로 사용자 고유의 3D 콘텐츠 저작이 가능하다. 자원의 재활용이 가능한 3D 객체와 모션등 대용량의 콘텐츠 자원을 활용한 시각적인 배치를 통해 장면을 손쉽게 완성할 수 있으며, 스크립트의 형태로 저장이 가능하다. 보다 편리한 저작을 위해 흐름도 기반으로 장면 구성이 가능하며, 영상인식기술 기반 인터페이스 지원을 통한 저작 환경을 제공한다[2].

컴퓨터 게임이나 3차원 애니메이션 등의 콘텐츠 제작에 있어서 객체의 움직임만이 필요한 것이 아니라, 움직임과 상황에 맞추어 음향적인 요소를 추가할 필요가 있다. 객체의 움직임이나 배경과 함께 음향적인 요소가 포함된다면 좀 더 풍성한 내용을 표현할 수 있을 것 이다. 하지만 필요한 소리를 찾거나 사용자의 의도에 맞게 만들어내고 수정하는 작업들에는 음향편집과 관련된 음향신호에 관한 지식과 기술이 요구되기 때문에 일반인에게는 쉽지 않다. 객체의 움직임을 스케치와 같은 직관적인 인터페이스로 만들어내는 것처럼[3] 객체의 움직임과 상황에 맞추어 소리도 직관적으로 만들어 낼 수 있는 방법이 있다면 콘텐츠 제작이 보다 용이해 질 것 이다.

이 논문에서는 COSMOS를 이용한 컴퓨터 게임이나 3차원 애니메이션 등의 콘텐츠 제작에 있어서 일반 사용자가 쉽고 간편하게 필요한 소리를 첨가할 수 있는 직관적인 입력 방법을 제안한다. 소리를 첨가할 때 원하는 소리에 대응되는 의성어를 사용자가 음성을 통하여 입력하면, 시스템이 자동으로 그 음향이 어떤 것인지 인식하여 미리 저장되어 있는 데이터베이스로부터 해당되는 음향샘플을 찾아낸다. 이후 사용자의 발성을 통하여 표현된 스타일로, 즉 의성어 발성의 시간적 길이와 음량에 맞추어 음향샘플을 변환한다. 이러한 방법을 이용하면 단순한 의성어 발성만으로 제작자의 의도에 맞게 음향을 검색 및 변환을 할 수 있게 된다 [4].

2. 관련연구

최근 콘텐츠 제작관련 저작도구의 사용자 인터페이스에 쉽고 간편하면서 직관적인 방법을 적용하기 위한 연구가 이어져 오고 있다. 특히 음성을 이용한 사용자 인터페이스 방법과 관련하여 연구된 예가 있다. Z. Wang과 M. Panne은 "Walk to here" 라는 시스템을 제안하고 구현하였다. 이 시스템은 음성명령을 이용하여 마치 감독이 영화를 찍을 때 연출을 지시하듯이 캐릭터의 움직임이나 카메라의 위치 및 각도, 대사입력 등을 제어 하였다. 이 논문에서는 애니메이션을 제작할 때의 모든 필요한 요소들을 모두 음성명령으로 제어 할 수 있어서 초보자들이 쉽게

* 교신저자

** 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0004938)

이용할 수 있다 [5]. 그러나 명령어로 이용되는 단어와 문법이 고정되어 있고, 사용자가 미리 이용방법에 대한 숙지가 필요하다는 단점이 있다.

T. Nakano 등 4명의 연구자들은 "Voice Drummer" 라는 시스템을 제안하고 구현하였는데, 이는 타악기의 음을 모사하는 음성을 입력방법으로 드럼연주에 대한 악보를 만들어내는 것이다. 이 시스템은 아무런 지식을 갖고 있지 않는 일반 사용자가 직관적으로 단순히 음성으로 음을 입력함으로써 드럼연주를 위한 작곡, 연습, 게임 등을 즐길 수 있다는 장점을 가지고 있다 [6].

위에서 본 것과 같이 지금까지의 연구는 주로 단순한 음성명령이나 악기소리를 표현하는 음성을 인식하는 것이었다. 하지만 본 논문에서는 컴퓨터 게임이나 3차원 애니메이션 등에서 상황의 표현을 위해 사용되는 음향신호를 만들어 내는 방법에 있어서 최대한 사용자가 쉽고 간편하게 사용할 수 있도록 직관적인 인터페이스를 활용하였다는 데에 차별성이 있다.

3. 연구내용

본 논문에서는 3D 콘텐츠 저작시스템에서 음향에 대한 오디오 신호를 의성어의 발성이라는 직관적인 방법을 통해 선택과 변환하는 방법을 제안한다.

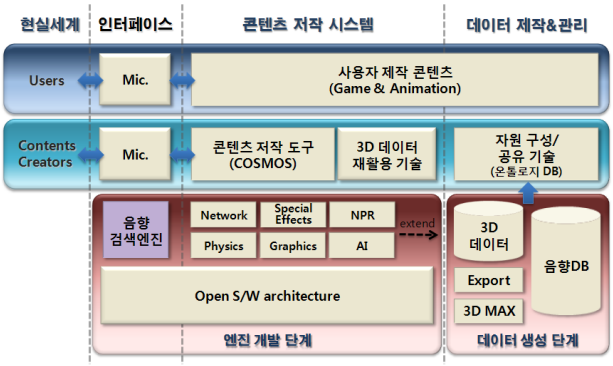

그림 1. 3D 콘텐츠 저작 시스템 개요

그림 1은 3D 콘텐츠 저작 시스템에서 음향 검색 엔진 기술을 도입한 그림으로, 이를 통해 사용자에게 편리한 음향 신호 저작 기능을 제공함으로써 제작의 효율성을 높일 수 있다. 음향검색엔진을 기반으로 사람의 언어로 소리를 흉내 내는 의성어 발성을 이용하여 원하는 음향샘플을 검색하고 이를 콘텐츠 제작 의도에 맞게 변환시켜 준다.

그림 2는 제작자가 의성어를 말하면 의성어와 대응되는 실제 음향 신호가 의성어 발성 형태에 따라 콘텐츠에 삽입되는 방법의 일례를 표현한 것이다. 골프게임을 제작할 때 필요한 각종 음향이나 애니메이션 등에서의 시계 알람소리, 자동차 충돌 소리 등을 첨가할 때, 의성어의 발성만으로도 주어진 샘플을 제작자 의도에 맞추어 변형시켜 주는 것이다.

그림 2. 직관적 음향정보 자동등록의 실시 예

COSMOS에 포함되는 음향정보 자동등록에 관한 블록도는 그림 3에서 볼 수 있는데, 제작자가 의성어를 음성으로 표현하면, 표현된 음향이 어떤 것인지 오디오 신호 분석을 통해 자동으로 미리 선정된 음향샘플들이 저장되어 있는 데이터 뱅크에서 찾아내게 된다. 사용자가 의성어를 음성으로 표현하면, 발성된 음성신호와 저장되어 있는 음향신호들의 패턴 비교를 통해 일치하거나 가장 가까운 것을 찾는다. 이를 위한 방법에는 여러 가지가 있을 수 있다. 이들 중 한 가지 방법은 음성인식을 이용하는 것이다. 인식하고자 하는 데이터 뱅크에 저장되어 있는 음향샘플은 각각 대표하는 음향에 대한 의성어의 메타데이터를 가지고 있고, 사용자가 의성어를 발성하면 이는 음성인식을 통해 텍스트 정보로 바뀌고 이를 데이터 뱅크의 메타데이터들과 비교하여 원하는 음향샘플을 찾아낸다. 이 방법은 의성어의 음성신호가 사용자에 의해 입력되는 음성신호와 비교되기 때문에 비교적 인식률이 높을 수 있다. 하지만 모든 사용자에 대해 독립적인 인식을 할 수 있도록 화자독립시스템을 만들기 위해는 많은 사람들에 의해 발성된 의성어들의 음성신호를 모아서 미리 특징벡터 수집을 해 놓아야 하는 어려움이 있다.

위의 어려움을 극복하기 위한 방법 중의 하나로 Generalized Likelihood Ratio(GLR) Test를 이용해 볼 수 있다[7]. 앞선 방법에서 수십 개 이상의 샘플데이터를 이용하여 모델을 만들었던 것과는 달리 한 번의 발성으로 얻어지는 데이터만을 이용하여서 패턴인식 과정을 수행할 수 있다는 데에 의의가 있다. 게다가 사용자가 시스템에서 지원하지 않는 새로운 음향을 추가하려고 할 때, 이에 대응되는 음성모델을 필요로 하지 않기 때문에 유용하게 쓰일 수 있다. 다만 사용자 독립적으로 사용할 때는 인식률이 하락할 가능성도 있어 사용자에 개인화 된 방식으로만 사용가능하다는 것이 약점이다.

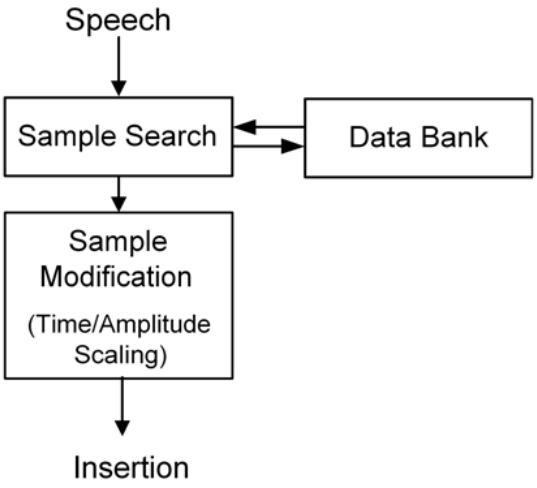

그림 3. 음성 인터페이스 기반 음향스케치 블록도

선택된 음향샘플 신호를 그대로 삽입하는 것이 아니라 제작자의 의도에 맞게 변형을 시킬 필요가 있다. 음향 샘플신호를 저작자의 발성행태에 따른 표현의도에 맞추어 시간 스케일(Time Scale)과 음량 등을 자동 변환하여, 최종적인 음향신호를 콘텐츠에 삽입해 준다. 특정 음향에 대한 하나의 샘플만 있어도 음성을 이용한 직관적인 인터페이스를 사용함으로써 제작자가 의성어를 발성하는 패턴을 조절하여 의도에 맞게 자유롭고 다양한 표현을 할 수 있다는 장점을 갖고 있다.

사용자의 의성어 발성신호를 기반으로 음향샘플 신호를 변형 시키는데 있어서 가장 중요한 기본단위는 음절이라고 할 수 있다. 음향신호가 음절 단위로 구분되지 않을지라도, 만약 사람이 인지하기에 두 음절로 여겨진다면, 의성어는 두 음절로 구성될 것이다. 그래서 각 음절마다의 경계를 파악하는 것이 우선되어야 하는데, 음절들 사이에 묵음구간이 있다면 이를 이용하여 각 음절마다의 시작점과 끝점을 찾을 수 있다. 하지만 의성어 발성 시에 음절과 음절사이에 묵음구간이 없다면 경계를 찾기가 매우 어렵다. 이러한 경우 음성신호에 있어서 각 음절마다 에너지의 중심은 모음이 발성되는 위치라고 생각해 볼 수 있고, 각 음절마다 모음이 한번 발성된다면 피크는 음절 당 하나가 존재한다고 가정할 수 있다. 그래서 음절 마다 하나의 피크를 찾고, 피크들 사이에서 최소의 에너지 값을 갖는 위치를 음절 간의 경계로 간주하였다.

시간적 스케일은 측정된 비율에 따라 Synchronized Overlap-Add (SOLA) Algorithm이라는 기본적인 방법을 이용하여 변환시킬 수 있다. 먼저 고정 길이의 겹쳐진 윈도우를 통해 일정간격(S_A)으로 입력된 신호(x)를 잘라낸다 [8].

$$x_m[n] = \begin{cases} x[mS_A + n] & \text{for } n = 0, \ldots, W-1 \\ 0 & otherwise \end{cases} \quad (1)$$

이후 윈도우는 원하는 비율에 따라 압축 또는 확장변환 되어 다시 합쳐지게 된다. 입력된 신호가 잘려지는 간격(S_A)과 합쳐지는 간격(S_S)의 비인 $\alpha = S_S / S_A$ 에 따라 신호가 변환되는데, $\alpha > 1$ 경우에는 확장 변환되고, $\alpha < 1$ 경우에는 압축 변환된다. 합쳐지는 단계(Overlapped-Add)에서 겹쳐서 비율이 변환된 신호들이 합쳐지는데, 이때 겹치는 부분이 최대한 유사하도록 간격을 미세조정 간격이 추가된다.

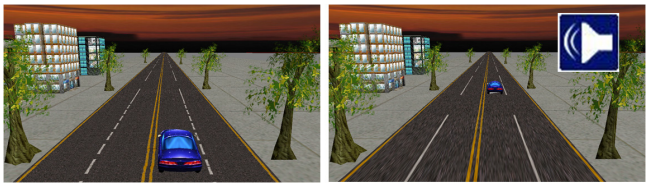

그림 4. 음향정보 자동 입력 기반 3D 콘텐츠 저작 장면

에너지 레벨의 변환은 음절 별로 의성어와 음향샘플의 에너지를(E_S , E_A)측정하여 각각 대응되는 음절 간의 에너지 비($\alpha = E_S / E_A$)를 계산하여 음향샘플의 샘플 값의 크기를 조절하였다.

$$x_o[n] = \alpha \bullet x_i[n] \text{ , for } n = 0, \ldots, N \quad (2)$$

여기에서 x_i 는 입력신호의 샘플이고, x_o 는 변환결과 신호의 샘플이다. N은 음절의 총 샘플 수를 의미한다.

4. COSMOS에서의 콘텐츠 제작 적용 결과

음향정보 자동입력 기능이 추가된 콘텐츠 저작 도구 COSMOS에서 콘텐츠를 제작해 보았으며, 그림 4는 자동차가 지나가는 콘텐츠 저작의 한 장면이다. 사용자가 자동차가 지나가는 장면에 맞춰 마이크를 통해 '부우웅'이라는 의성어를 입력하였고, 저작 장면에 자동으로 사운드가 변환되어 자동차 효과음이 삽입됨을 확인하였다. 이 밖에 자동차 경적소리, 동물은 울음소리, 초인종 소리 등 다양한 음향정보에 대한 적용을 실험 중에 있다.

참고문헌

[1] Su Mi Ji, Sung Wook Baik, "A Study on Contents Manufacturing System for Massive Contents Production", 한국멀티미디어학회, Vol.13, No.12, pp.1832-1842, 2010

[2] 지수미, 이정중, 김성국, 우경덕, 백성욱, "효율적인 3D 게임 및 애니메이션 콘텐츠 제작을 위한 직관적인 저작 기술 개발", 한국멀티미디어학회, Vol.13, No.5, pp.780-791, 2010

[3] Collomosse, J. P., McNeill, G., Qian, Y., "Storyboard sketches for Content Based Video Retrieval", Computer Vision, 2009 IEEE 12th International Conference on, pp.245 - 252, 2009

[4] S. Kwon and L. H. Kim, "Sound Sketching Via Voice," ACM International Conference on Ubiquitous Information Management and Communication 2011 (ICUIMC 2011), Febraury 21-23, 2011.

[5] T. Nakano, M. Goto, J. Ogata, and Y. Hiraga, "Voice Drummer: A Music Notation Interface of Drum Sounds Using Voice Percussion Input," In Proc. of ACM Symposium on User Interface Software and Technology, p. 49-50, 2005.

[6] Z. Wang and M. Panne, "Walk to here: A Voice Driven Animation System," In Proc. of Eurographics/ ACM SIGGRAPH Symposium on Computer Animation, p. 16-20, 2006.

[7] S. Kwon and S. Narayanan, "Unsupervised Speaker Indexing Using Generic Models," IEEE Transactions On Speech and Audio Processing, p. 1004-1013, 2005

[8] D. Hejna and B. Musicus, "The SOLAFS Time-Scale Modification Algorithm," Technical report of BBN, 1991.