애니메이션 저작도구를 위한 음성 기반 음향 스케치

Voice Driven Sound Sketch for Animation Authoring Tools

권순일
세종대학교 컴퓨터공학부 디지털콘텐츠전공

Soon-Il Kwon(sikwon@sejong.ac.kr)

요약

애니메이션 캐릭터의 동작을 만들어내기 위해 펜으로 스케치하는 형식의 인터페이스를 이용하는 저작도구들이 연구되어 왔지만, 아직까지 음향적인 요소에 있어서 직관적인 인터페이스를 사용하여 만들어내는 방법은 연구되지 않았다. 본 논문에서는 사용자가 음향과 대응되는 의성어의 발성을 통하여 표현하면 이에 대응되는 음향샘플이 선택되어 삽입되는 방법을 제안하고자 한다. 일반적으로 사용되는 통계적 모델을 기반으로 하는 패턴인식 방법을 이용하여 의성어 발성만으로 대응되는 음향샘플을 어느 정도 인식할 수 있는지를 실험해본 결과 의성어의 음성샘플을 이용한 경우 최대 97%의 인식률을 얻을 수 있었다. 또한 새로운 음향샘플 등록 시에 발생하는 음성데이터 수집의 어려움을 극복하기 위하여 음성모델을 만드는 대신에 의성어의 음성샘플 하나만 사용하는 GLR Test를 활용해보니 기존의 방법과 거의 대등한 인식률을 실험적으로 확인할 수 있었다.

■ 중심어 : | 애니메이션 저작도구 | 음향 스케치 | 오디오 사용자 인터페이스 | 컴퓨터 사용자 상호작용 | 음성 인터페이스 | 의성어 |

Abstract

Authoring tools for sketching the motion of characters to be animated have been studied. However the natural interface for sound editing has not been sufficiently studied. In this paper, I present a novel method that sound sample is selected by speaking sound-imitation words(onomatopoeia). Experiment with the method based on statistical models, which is generally used for pattern recognition, showed up to 97% in the accuracy of recognition. In addition, to address the difficulty of data collection for newly enrolled sound samples, the GLR Test based on only one sample of each sound-imitation word showed almost the same accuracy as the previous method.

■ keyword : | Animation Authoring Tool | Sound Sketch | Audio User Interface | Human-Computer Interaction | Speech Interface | Onomatopoeia |

I. 서 론

최근 애니메이션에 대한 관심은 단순히 예술작품이나 상업적인 목적의 영상물에서 끝나지 않고, 일반 대중들의 직접적인 제작에 대한 기회까지 가져다주고 있다. 애니메이션 제작에 있어서 대중들이 직접 참여하게

* 이 논문은 2009년도 세종대학교 교내연구비 지원에 의한 논문입니다.
접수번호 : #091210-002
접수일자 : 2009년 12월 10일
심사완료일 : 2010년 02월 11일
교신저자 : 권순일, e-mail : sikwon@sejong.ac.kr

되면서 다양한 내용과 수준의 애니메이션 콘텐츠, 즉 일반인들이 각자의 취향이나 의도에 맞는 콘텐츠가 양산되고 있다. 하지만 전문적으로 애니메이션 제작을 하는 사람과 달리 일반인들이 사용할 수 있는 도구가 제한적이고, 제공된다고 해도 애니메이션 제작에 대한 경험 부족과 도구에 대한 사용 미숙으로 어려움을 겪고 있다. 이러한 이유로 일반 사용자가 보다 쉽고 간편하게 원하는 내용의 애니메이션을 제작할 수 있도록 만들어진 애니메이션 저작 도구가 요구되고 있다 [1-3][10].

애니메이션을 제작하는데 있어서 등장 캐릭터의 동작들과 음향효과, 대화음성 등 많은 작업들이 필요하다. 이러한 작업들을 수행하기 위해 애니메이션 제작에 필요한 도구를 사용한다고 하더라도 전문가와는 달리 일반인은 작업 하나하나에 대한 전문지식과 기술이 부족하다. 그래서 일반인들이 기존의 전문가 용 애니메이션 저작도구를 사용하기 위해서는 많은 시간과 노력이 요구된다. 이러한 문제를 해결하기 위해 직관적인 인터페이스 방법을 활용하여 애니메이션 캐릭터의 동작들을 만들어내는 도구들이 연구개발 되고 있다. 일반적으로 펜과 종이 또는 분필과 칠판을 이용하여 무언가를 그리듯이 캐릭터의 움직임을 스케치할 수 있다. 즉 스케치라는 입력방법을 통해 사용자는 자유롭게 그림을 그리듯이 캐릭터 동작을 만들어 낼 수 있다. 그런데 캐릭터의 동작만이 애니메이션을 제작하는데 필요한 것이 아니라 동작과 상황에 맞추어 음향적인 요소를 첨가할 필요가 있다. 캐릭터의 동작을 만들어내는 것과 별도로 필요한 소리를 찾거나 만들어 내고 제작자의 의도에 맞게 수정하는 작업이 뒤따르게 되는데, 이 또한 음향에 관한 지식과 기술이 요구된다 [4-7][11].

이 논문에서는 애니메이션 제작에 있어서 일반 사용자가 쉽고 간편하게 필요한 음향을 애니메이션에 첨가할 수 있도록 직관적인 입력 방법을 제안하고자 한다. 구체적으로 말하자면, 캐릭터의 동작과 상황에 맞추어 음향을 첨가할 때 원하는 음향에 대응되는 의성어를 사용자가 음성을 통하여 표현하면, 시스템이 자동으로 그 음향이 어떤 것인지 인식하여 미리 저장되어 있는 데이터베이스로부터 해당되는 음향샘플을 찾아내고, 사용자가 원하는 시간적 위치에 음향샘플을 첨가하게 된다. 이러한 방법을 이용하면 단순한 의성어 발성만으로 제작자의 의도한 음향을 첨가할 수 있게 된다.

본 논문은 다음과 같이 구성되어 있다. 2장 관련연구에서는 직관적인 인터페이스 방법을 이용하는 애니메이션 저작방법과 관련된 주요연구들에 대해 소개하고 이번 논문의 연구와 비교한다. 3장 제안내용에서는 이번 논문에서 제안하는 의성어를 발성하여 애니메이션에 음향을 스케치하는 방법에 대해 전체적인 모습을 설명하고, 의성어 발성에 의한 음향데이터 인식을 위해 사용된 방법에 대해 구체적으로 설명한 후, 4장 실험에서 위에 따른 실험결과를 제시하고 비교분석한다. 마지막으로 본 논문의 결론을 내린다.

II. 관련연구

최근 10년간 애니메이션 제작 관련 저작도구의 사용자 인터페이스에 쉽고 간편하면서 직관적인 방법을 적용하기 위한 연구가 이어져 오고 있다. 애니메이션의 캐릭터들의 동작을 손쉽게 구현할 수 있도록 하기 위해 M. Thorne, D. Burke, 그리고 M. Panne은 "Motion Doodles" 라는 시스템을 제안하고 구현하였다. 캐릭터의 동작들을 만들 때, 펜으로 선(Lines)이나, 호(Arcs), 고리(Loops) 등을 그림 그리듯이 입력함으로써 그에 대응되는 일련의 동작이 생성된다. 기본이 되는 18개의 단위 이차원적인 동작들(2D Motions)과 이에 일부에 해당하는 삼차원적인 동작들(3D Motions)이 미리 저장되어 있고, 입력 단에서 단순하게 그려지는 그림을 데이터베이스에 있는 동작들을 엮어서 만들어 내기 때문에 일반 사용자와 같은 초보자가 쉽게 애니메이션을 제작할 수 있다[1]. 다양한 동작에 대한 데이터베이스를 구축하는 것에 시간과 노력이 요구되고 있고 아직까지는 일부 동작에 대해서만 준비가 되어 있어서 제한된 동작들에 대해서만 이용할 수 있다는 단점이 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 제어할 수 있게 하였다. 이 논문에서는 애니메이션을 제작할 때의 모든 필요한 요소들을 모두 음성명령으로 제어 할 수 있어서 초보자들이 쉽게 이용할 수 있다[3]. 하지만 명령어로 이용되는 단어와 문법이 고정되어 있어서 어느 정도의 이용방법에 대한 숙지가 필요하다는 단점이 여전히 남아있다. 대사에 대한 입력은 동작과 대응되는 시간적 위치에 맞추어 사용자가 한 말을 저장해 놓았다가 실행 시에 재생하는 방법을 사용하고 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[2]. 하지만 아직까지는 타악기의 연주라는 영역에 한정되어 있어서 추가적인 연구로 그 영역을 넓힐 필요가 있어 보인다.

위에서 본 것과 같이 지금까지의 연구는 주로 애니메이션 캐릭터는 동작을 만들어내는 데 있어서 직관적인 인터페이스 방법에 대한 연구가 주를 이루고 있다. 인터페이스 방법에 있어서는 주로 펜이나 손가락으로 선을 긋는 동작인식 또는 음성명령을 인식하는 것 들이 최근까지 연구되었다. 하지만 본 논문에서는 캐릭터의 동작이 아닌 동작이나 상황의 표현을 위해 사용되는 음향신호를 찾는데 있어서 최대한 사용자가 쉽고 간편하게 사용할 수 있도록 직관적인 인터페이스를 활용하였다는 데에 차별성이 있다.

III. 제안내용

1. 의성어와 음향신호

의성어는 국어사전에 의하면 넓은 의미에서 언어 기호의 음성형식과 그 의미 내용과의 사이에 필연적인 관계가 성립하는 일군의 단어를 총칭하며, 좁은 의미로는 사물의 소리를 흉내를 낸 말이다. 의성어는 자연적 또는 인공적인 모든 소리를 지칭하거나 묘사하기 위해, 최대한 그 소리에 가까우면서도 해당 언어의 음운과 음절 구조에 맞도록 만들어진 말이다. 소리를 인간의 말로 바꾸어 들려주는 것이므로, 인간의 청각 기관과 발음기관이 같은 한 여러 언어 간에 의성어가 상당히 비슷해질 가능성이 있다.

의성어의 음성신호가 그 대상이 된 음향신호에 가까운 소리를 표현하는 것이지만, 완전히 똑같은 신호를

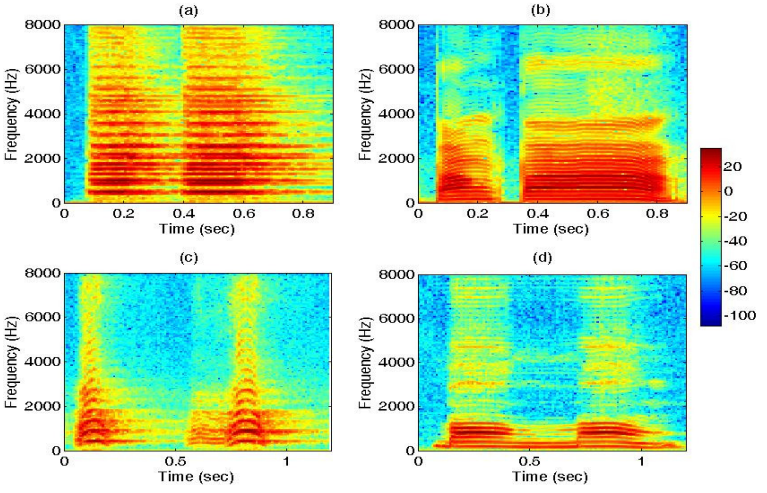

그림 1. 실제 음향신호와 의성어 음성신호의 스펙트로그램 비교: (a) 자동차 경적소리의 음향신호, (b) '빵빵'의 음성신호, (c) 개 짖는 소리의 음향신호, (d) '멍멍'의 음성신호

만들어 내지는 못할 것으로 생각되었다. 그래서 음향신호와 음향신호를 흉내 낸 의성어의 음성신호를 분석해 보았다. [그림 1]에서는 두 가지의 실제 음향신호와 이에 상응하는 의성어의 음성신호를 보여주고 있다. [그림 1](a)는 자동차 경적소리에 해당되는 음향신호의 스펙트로그램이고, [그림 1](b)는 자동차 경적소리를 표현하는 의성어인 '빵빵'의 음성신호 스펙트로그램이다. 그림에서 볼 수 있듯이 실제 음향신호는 기계적으로 만들어진 인공적인 소리인데다가 전 주파수 대역에 걸쳐서 비교적 고르게 에너지가 분포되어 있지만, 의성어의 음성신호는 사람이 만들어 낸 소리이기 때문에 약 4000Hz 이하의 주파수 대역에 대부분의 에너지가 집중되어 있다. [그림 1](c)는 개가 짖는 소리에 해당되는 실제 음향신호의 스펙트로그램이고, [그림 1](d)는 개가 짖는 소리의 의성어인 '멍멍'의 음성신호 스펙트로그램이다. 그림에서 볼 수 있듯이 실제 음향신호는 동물이 낸 소리이기 때문에 일정한 주파수 대역에 에너지의 비중이 높은 것을 볼 수 있고, 음성신호에서 볼 수 있는 포먼트(Formant) 같은 띠들이 보이지만, 전체적으로 음성신호와는 다른 패턴을 가지고 있다. 이는 사람과 개의 발성기관이 서로 상이하기 때문일 것으로 추정된다.

2. 음성기반 음향 스케치

사물이나 동물이 만들어내는 소리는 애니메이션 제작에 있어서 음향적 요소로 큰 역할을 한다. 하지만, 이러한 음향들을 사용하기 위해서는 먼저 음향 샘플이 필요하고, 샘플이 있다면 이를 원하는 위치에 삽입을 해야 한다. 샘플을 구하는 것은 직접 녹음을 하거나 상용화되어있는 데이터 뱅크를 이용할 수 있지만, 기본 샘플을 저작자의 의도에 따라 삽입 또는 변형시키는 것은 저작자의 몫이 된다. 일반적으로 애니메이션 저작자들이 음향신호의 조작에 익숙지 않다면, 샘플이 있어도 의도에 따라 이용하기가 어려워 처음부터 음향적 요소를 배제하는 경우가 많다. 애니메이션을 풍성하게 만들어주는 음향을 저작자들이 쉽게 사용 할 수 있게 해 주는 방법이 요구되고 있다.

본 논문에서는 애니메이션 제작에 있어서 음향에 대한 오디오 신호를 의성어의 발성이라는 직관적인 방법을 통해 원하는 음향샘플을 검색하는 방법을 제안한다. [그림 2]는 저작자가 의성어를 말하면 이를 의성어와 대응되는 실제 음향 신호를 애니메이션에 삽입시켜주는 방법에 대한 예를 표현한 것이다. 즉 저작자가 자동차 간의 운행 중에 발생할 수 있는 경적소리('빵빵')나 급정거하는 소리('끼익')를 말하면, 이에 해당되는 실제 음향을 저작자가 원하는 시간에 맞추어 삽입해 주는 것이다.

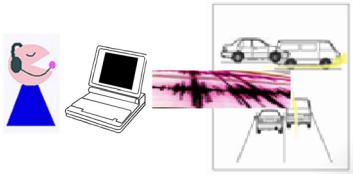

그림 2. 애니메이션 저작을 위한 음성 기반 음향스케치 예시

전체적인 음향스케치 방법의 블록도는 [그림 3]에서 볼 수 있는데, 저작자가 의성어를 음성으로 표현하면 그 음향이 어떤 것인지 오디오 신호 분석을 통해 자동으로 인식한 후, 미리 저장되어 있는 데이터 뱅크에서 해당되는 오디오 신호 샘플을 검색하게 되고, 그 결과 검색된 음향 샘플신호를 애니메이션의 오디오 트랙에 삽입해 준다.

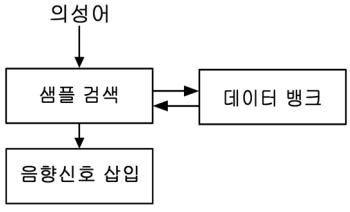

그림 3. 음성기반 음향스케치 블록도

3. 음성신호에 의한 음향신호 인식

음성기반 음향스케치에서는 저작자가 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 인식하고자 하는 데이터 뱅

크에 저장되어 있는 음향샘플과 대응되는 의성어에 대해서 특징벡터(Feature Vector)를 미리 추출하여 놓고, 이 후 저작자에 의해 발성되는 의성어의 음성신호로부터 특성벡터를 추출하여 데이터 뱅크에 있는 의성어들의 특징벡터와의 패턴 비교를 통해 인식하는 과정을 거쳐 음향샘플을 검색해 내는 방법이 있다[그림 4]. 이는 기존의 음성신호와 관련된 패턴인식의 전형적인 방법이다. 이 방법은 의성어의 음성신호가 저작자에 의해 입력되는 음성신호와 비교되기 때문에 비교적 인식률이 높을 수 있다. 하지만 모든 저작자에 대해 독립적인 인식을 할 수 있도록 화자독립시스템을 만들기 위해 많은 사람들에 의해 발성된 의성어들의 음성신호를 모아서 미리 특징벡터 수집을 해 놓아야 하는 어려움이 있다.

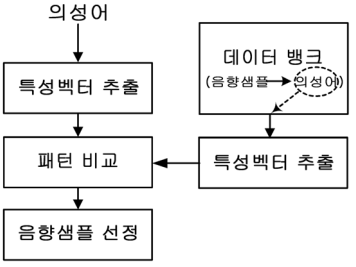

그림 4. 음성신호 간의 패턴비교 방법

위의 어려움을 극복하기 위한 방법으로는 음향샘플 신호의 특징벡터를 미리 추출하여 데이터 뱅크에 가지고 있고, 저작자에 의해 발성되는 음성신호로부터 특징벡터를 추출하여 데이터 뱅크에 있는 음향샘플의 특징벡터와의 패턴비교 방법을 통해 인식해 내는 방법이 있다[그림 5]. 저작자에 독립적인 음성신호의 패턴을 저장해 놓을 필요가 없고, 대표되는 음향의 패턴만을 보관하면 되기 때문에 앞선 방법들의 여러 사람에 대한 음성데이터 수집이라는 시간과 노력, 금전적 어려움은 피할 수 있지만, 서로 상이한 특성을 가지는 음성과 음향을 비교하는 것이기 때문에 음성과 음성을 비교하는 것보다 인식 성능이 하락할 수 있다는 우려가 있다.

위의 어려움을 극복하기 위한 또 하나의 방법으로 다양한 사람들의 음성샘플을 수집하여 의성어 별 모델을 만드는 것이 어려울 경우에 한해서 이를 대체할 방법을 생각해 보았다. 만약 사용자가 시스템에서 지원하지 않은 새로운 음향을 추가하려고 할 때, 이에 대응되는 음성모델을 어떻게 만들 것인가가 문제가 될 것이다. 일단 사용을 시작하면 특정 사용자만이 저작도구를 사용할 것이라는 가정을 하고, 개인화 된 의성어 인식 방법으로 모델이 필요하지 않은 좀 더 편리한 개인화 된 의성어 인식 방법을 생각해 보았다. Generalized Likelihood Ratio(GLR) Test를 이용하여 패턴을 인식할 경우 통계적 모델을 만들어야 하는 번거로움을 피할 수 있다. 앞선 방법에서 수십 개 이상의 샘플데이터를 이용하여 GMM이라는 모델을 만들었던 것과는 달리 한 번의 발성으로 얻어지는 데이터만을 이용하여서 패턴인식과정을 수행할 수 있다는 데에 의의가 있다. 다만 사용자 독립적으로 사용하기는 힘들고 인식률이 하락할 가능성도 있다[9].

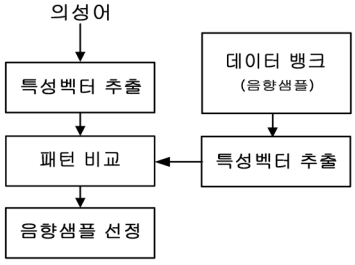

그림 5. 음성신호와 음향샘플 간의 패턴비교 방법

IV. 실험

본 논문에서는 앞에서 소개된 방법들에 의한 음향샘플의 인식에 대한 성능 평가를 위해 실험을 해 보았다. 애니메이션에 사용되는 음향의 종류는 매우 많을 것이지만, 그 중에서 가장 많이 사용될 만한 음향을 4가지정도 선정해 보았다. 동물의 울음소리를 대표하여 개가 짖는 소리, 기계음과 관련하여 현관문 벨소리, 자동차 경적소리, 그리고 사람에 의해 만들어지는 기침 또는 재채기 소리를 실험의 대상으로 삼았다. 또한 각 음향에 해당하는 의성어들, 즉 '멍멍', '딩동', '빵빵', 그리고 '에취'를 발성한 음성신호를 수집하였다. 의성어의 음성 데이터는 20대 남성 35명과 여성 15명, 총 50명에 의

해 각 의성어 별 한 개씩 만들어졌고, 음향샘플 신호는 게임재료파크(www.kgdb.or.kr)의 음향 샘플을 이용하였다. 모든 오디오 신호는 16khz의 샘플링 주파수와 16bit, 모노채널 샘플이었다.

1. 음성모델을 이용한 방법

음성신호의 패턴을 확률통계적인 모델을 만들어 이를 기반으로 의성어를 인식하는 실험을 해 보았다. 테스트 이전에 미리 실험 대상이 되는 의성어에 대해서 특징벡터를 추출하였다. 특징벡터는 Mel-Frequency Cepstral Coefficient(MFCC) 라는 음성인식에서 주로 사용되는 특징벡터를 사용하였다. 각 의성어 별로 특징벡터들을 이용하여 Gaussian Mixture Model (GMM) 이라는 확률통계적인 모델을 만들었다. Mixture의 개수는 최적의 수를 미리 예측하기 어렵기 때문에 8개와 16개의 두 가지로 각각 만들어 실험해 보았다. 이렇게 만들어진 각 의성어 별 음성모델을 기반으로 입력 단으로부터 의성어 음성신호가 들어오면 특징벡터를 추출하여 확률적으로 판단하여 인식률을 계산하였다. 실험은 Cross Validation 방법으로 반복적으로 수행하였고, 그 결과를 [표 1]과 [표 2]의 혼동행렬(Confusion Matrix)로 나타내었다[8].

표 1. Mixture 수가 8 일 때의 음성모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**88**	2	0	10
	딩동	2	**88**	2	8
	에취	0	0	**100**	0
	멍멍	2	0	0	**98**

표 2. Mixture 수가 16 일 때의 음성모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**96**	0	0	4
	딩동	2	**96**	0	2
	에취	0	0	**100**	0
	멍멍	2	2	0	**96**

[표 1]과 [표 2]의 결과에서는 공통적으로 상당히 높은 인식률을 보였다. '빵빵'과 '멍멍'이 다소 혼동되기도 하기만, Mixture 수를 늘렸을 때, 인식 오류는 상당히 줄어드는 모습을 보였다. 인식 대상이 되는 의성어의 음성을 화자독립이 될 수 있도록 다양한 사람들의 음성으로 모델을 만들었기 때문에, 한번 만들어진 모델을 사용자 누구에게나 적용될 수 있다는 장점이 있다. 하지만, 의성어 하나하나마다 많은 사람들의 음성샘플 데이터를 수집하여 각각의 모델을 만들어야 한다는 어려움이 있다.

2. 음향모델을 이용한 방법

두 번째 방법에서는 사용하고자 하는 음향샘플 신호의 특징벡터를 미리 추출하여 데이터 뱅크에 음향샘플들의 GMM 모델을 훈련시켜 이들을 미리 저장해 놓았다. 모델을 만들기 위해 각 음향샘플 별로 50개의 데이터를 수집하여 사용하였다. 특징벡터는 앞선 방법에서 사용한 것과 동일하다. 실험의 테스트에서는 저작자에 의해 발성되는 음성신호로부터 특징벡터를 추출하여 데이터 뱅크에 있는 음향샘플의 특징벡터와 확률적 계산을 통해 인식률을 계산했다. 실험은 Cross Validation 방법으로 반복적으로 수행하였고, 그 결과를 [표 3]과 [표 4]의 혼동행렬로 나타내었다.

표 3. Mixture 수가 8 일 때의 음향모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**4**	14	22	60
	딩동	0	**10**	66	24
	에취	0	0	**100**	0
	멍멍	0	18	28	**54**

표 4. Mixture 수가 16 일 때의 음향모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**0**	2	36	62
	딩동	0	**0**	50	50
	에취	0	0	**96**	4
	멍멍	0	0	26	**74**

[표 3]과 [표 4]에서 모두 매우 낮은 인식률을 보여주고 있다. 네 가지 음향 중에서 '에취'는 음성모델을 사용한 방법과 대등한 높은 인식률을 보여주었고, '멍멍'은 조금 낮은 인식률을 보여주었다. 하지만 나머지 음향들은 너무 낮은 인식률을 보여 주었다. '에취'라는 음향은 사람의 재채기 소리이기 때문에 음성이 아니더라도 다른 음향샘플에 비하면 거의 음성에 가깝기 때문에 인식률이 좋았다고 생각된다. 이 실험 결과로 볼 때, 기존의 특징벡터(MFCC)외에 의성어와 음향을 연결 지을 수 있는 새로운 특징추출이 필요한 것으로 판단되었다.

3. 화자종속적 방법

화자종속 방법에 있어서 화자종속 모델을 만들어 사용하는 방법은 기존에 있어왔던 것이다. 비교를 위해 화자종속 모델기반 방법도 실험해 보았다. 테스트 이전에 미리 실험 대상이 되는 특정인 한 명의 의성어에 대해서 특징벡터를 추출하여 모델을 만들었다. 이렇게 만들어진 각 의성어 별 모델을 기반으로 입력 단으로부터 의성어 음성신호가 들어오면 특징벡터를 추출하여 확률적으로 판단하여 인식률을 계산하였다. [표 5]와 [표 6]의 실험결과에서는 모두 상당히 높은 인식률을 보였다. 음성인식 및 화자인식이라는 측면에서 개인화 된 모델을 이용할 경우인데다가, 본 논문에서는 대표적인 4개의 음향만을 실험했기 때문에 높은 인식률을 보이는 것은 한편으로 당연한 것이다. 하지만 이 방법 또한 특정인에 있어서 많은 음향샘플 별 의성어 음성 데이터를 수집해야하는 어려움은 여전히 남아 있다.

표 5. Mixture 수가 8 일 때의 개인화된 음성모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**100**	0	0	0
	딩동	0	**100**	0	0
	에취	0	0	**100**	0
	멍멍	0	0	0	**100**

표 6. Mixture 수가 16 일 때의 개인화된 음성모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**100**	0	0	0
	딩동	0	**100**	0	0
	에취	0	0	**100**	0
	멍멍	0	0	0	**100**

위와 같은 기존방법과 비교하여 좀 더 편리한 개인화된 의성어 인식 방법을 만들어 보았다. Generalized Likelihood Ratio(GLR) Test를 이용하여 패턴을 인식해 보았다. 이전에 미리 실험 대상이 되는 특정인 한 명의 의성어에 대한 음성데이터를 한 개만 수집하여 음향샘플에 대응시켜놓고, 동일인이 의성어를 발성해서 네 가지 음향 중에서 어느 것과 일치하는 것으로 나오는지를 알아보았다. 그 결과인 [표 7]에서는 사용자의 개인화된 GMM모델을 만들어서 수행했던 것 보다는 조금 낮은 인식률을 보였지만, 상당히 높은 인식률을 유지할 수 있다는 것을 알 수 있었다. 이 실험 결과를 바탕으로 하여 시스템에 미리 들어가게 되는 음향샘플들은 화자독립의 음성모델을 만들어 놓고, 사용자에 의해 새롭게 추가되는 음향샘플에는 한 번만 사용자에게 의성어를 발성하게 한 후 이를 음향샘플과 대응시켜 놓으면, 전체적으로 우수한 인식률을 유지하면서도 사용자가 새로운 음향을 추가하는데 편리한 방법마련이 가능하리라 생각된다.

표 7. GLR Test를 이용한 개인화된 음성모델 기반 의성어 인식 결과 (단위, %)

		인식 결과			
		빵빵	딩동	에취	멍멍
실제 음성	빵빵	**100**	0	0	0
	딩동	0	**97**	1	2
	에취	0	0	**100**	0
	멍멍	0	0	0	**100**

V. 결 론

애니메이션 제작 시에 원하는 음향을 삽입하는데 있

어서 그 음향과 연관된 의성어의 발성만으로 음향샘플을 선택할 수 직관적인 인터페이스 방법을 만드는 것이 이번 연구의 목적이었다. 의성어를 통한 음향샘플 선택에 있어서 여러 가지 방법이 있을 수 있겠지만, 최대한 간편하면서도 정확도가 높고 사용자 편의성을 고려한 방법을 구상하였다. 이를 위해 의성어의 음성샘플들을 수집하여 화자독립적인 음성모델을 만드는 일반적인 방법과 음향샘플들을 수집하여 음향모델을 만드는 새로운 방법을 실험해 보았다. 그 결과 음향모델을 이용하여 의성어 발성만으로 이에 대응되는 음향샘플을 찾아내는데 한계를 보였다. 이를 해결하기 위해서는 기존의 특성벡터 이외에 새로운 특성벡터를 발굴해 낼 필요가 있다는 결론을 얻었다.

위의 방법들은 화자독립적인 음성모델을 만드는 방법으로 사용자가 직접 수십 명의 음성샘플을 수집하고 모델을 만들기는 쉽지 않다. 만일 임의의 사용자가 저작도구 사용 중에 새롭게 추가하고 싶은 음향샘플이 있을 때 문제가 될 수 있다. 이를 보완하기 위해 많은 음성데이터 수집을 피하는 방법으로 GLR Test를 이용하여 의성어 발성을 통한 음향샘플 검색을 해 본 결과 사용자 중심의 음성모델, 즉 화자종속 모델을 이용하는 기존의 방법과 대등한 인식률을 보였다.

본 논문에서 제시한 시스템을 이용한다면 애니메이션 저작 시에 단지 캐릭터의 동작 또는 간단한 배경그림만 만드는 것이 아니라 간단한 음향효과도 쉽고 편리하게 넣을 수 있다. 또한 이는 애니메이션 저작도구의 음성 또는 오디오 인터페이스의 새로운 시도이고 앞으로 꾸준히 연구해 볼 흥미로운 주제가 될 것이다.

참 고 문 헌

[1] M. Thorne, D. Burke, and M. Panne, "Motion Doodles: An Interface for Sketching Character Motion," ACM Transactions on Graphics, Vol. 23, pp.424-431, 2004.

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

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

[4] O. Gillet and G. Richard, "Indexing and Querying Drum Loops Databases," Proc. of International workshop on Content Based on Multimedia and Indexing (CBMI'05), Riga, Latvia, 2005(6).

[5] K. Ishihara, Y. Tsubota, and H. G. Okuno, "Automatic Transformation of Environmental Sounds into Sound-Imitation Words Baed on Japanese Syllable Structure," Proc. of European Conference on Speech Communication and Technology, pp.3185-3188, 2003.

[6] K. Ishihara, T. Nakatani, T. Ogata, and H. G. Okuno, "Automatic Sound-Imitation Word Recognition from Environmental Sounds Focusing on Ambiguity Problem in Determining Phonemes," Lecture Note on Artificial Intelligence, Vol.3157, pp.909-918, 2004.

[7] T. C. Andringa and M. E. Niessen, "Real-world sound recognition: A recipe," Proc. of the 1st Workshop on Learning Semantics in Audio Signals(LSAS 2006), pp.106-118, 2006.

[8] R. Duda, D. Stork, and P. Hart, *Pattern Classification*, Wiley-Interscience Pub, 2/E, 2000.

[9] J. S. Baek, "A Generalized Likelihood Ratio Test in Outlier Detection," Korean Journal of Applied Statistics, Vol.4, pp.225-237, 1994.

[10] R. C. Davis, B. Colwell, and J. A. Landay, “K-sketch: a ‘kinetic’ sketch pad for novice animators,” Proc. of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp.413-422, 2008.

[11] M. Battermann, S. Heise, and J. Loviscach, “SonoSketch: Querying Sound Effect Databases through Painting,” Proc. of 126th AES Convention, Paper Number 7794, 2009.

저 자 소 개

권 순 일(Soon-Il Kwon) 정회원

- 1998년 2월 : 연세대학교 전자공학과(공학사)
- 2000년 5월 : 미국 University of Southern California 전기공학과 졸업(공학석사)
- 2005년 5월 : 미국 University of Southern California 전기공학과 졸업(공학박사)
- 2005년 8월 ~ 2006년 7월 : 삼성전자 책임연구원
- 2006년 8월 ~ 2009년 2월 : 한국과학기술연구원 선임연구원
- 2009년 3월 ~ 현재 : 세종대학교 교수

<관심분야> : 음성인식, 화자인식, 음성/오디오 인터페이스, HCI, HRI