2003년도 대한전자공학회 하계종합학술대회 제26권 제1호

인공생명체의 감정표현을 위한 음성처리

장국현, 한동주, 이상훈*, 서일홍
한양대학교 정보통신대학원 정보통신공학과(Tel: 02-2290-0392; E-mail: ihsuh@hanyang.ac.kr)
*한양대학교 전자전기제어계측공학과(Tel: 031-400-3717; E-mail: shlee@incorl.hanyang.ac.kr)

Emotional Text-to-Speech System for Artificial Life Systems

Guo-Xuan Zhang, Dong-Ju Han, Sang-Hoon Lee*, Il-Hong Suh
The Graduate School of Information & Communications, Hanyang University
*School of Electrical Engineering and Computer Science, Hanyang University
E-mail: ihsuh@hanyang.ac.kr

Abstract

인간과 인공생명체(Artificial Life Systems)가 서로 커뮤니케이션을 진행하기 위하여 인공생명체는 자신이 의도한 바를 음성, 표정, 행동 등 다양한 방식을 통하여 표현할 수 있어야 한다. 특히 자신의 좋아함과 싫음 등 자율적인 감정을 표현할 수 있는 것은 인공생명체가 더욱 지능적이고 실제 생명체의 특성을 가지게 되는 중요한 전제조건이기도 하다.

위에서 언급한 인공생명체의 감정표현 특성을 구현하기 위하여 본 논문에서는 음성 속에 감정을 포함시키는 방법을 제안한다. 먼저 인간의 감정표현 음성데이터를 실제로 구축하고 이러한 음성데이터에서 감정을 표현하는데 사용되는 에너지, 지속시간, 피치(pitch) 등 특징을 추출한 후, 일반적인 음성에 위 과정에서 추출한 감정표현 특징을 적용하였으며 부가적인 주파수대역 필터링을 통해 기쁨, 슬픔, 화남, 두려움, 혐오, 놀람 등 6가지 감정을 표현할 수 있게 하였다. 감정표현을 위한 음성처리 알고리즘은 현재 음성합성에서 가장 널리 사용되고 있는 TD-PSOLA[1] 방법을 사용하였다.

I. 서론

음성을 통한 감정표현 과정은 음성합성과 같은 처리과정을 통하여 감정표현에 필요한 특징을 일반 음성에 추가함으로써 완성된다. 따라서 인공생명체의 음성을 통한 감정표현 과정을 구현하기 위해서는 먼저 음성합성 방식에 대한 사전 연구와 이러한 방식상의 차이로 인하여 발생하는 문제점을 분석해야 한다.

현재 실제로 사용되고 있는 음성합성방식에는 크게 포만트(formant) 합성방식과 결합(concatenative) 합성방식으로 나눌 수 있으며, 따라서 음성을 통한 감정표현도 위 두 가지 방법 중 어느 것을 택하느냐에 따라 많이 달라지게 된다. 포만트 합성방식은 음원-필터(source-filter)이론에 의하여 인간의 음성신호를 직접 생성하는 방식을 가리키며, 이러한 방식을 이용한 감정표현은 제어가 가능한 파라미터가 많기 때문에 표현된 감정의 품질이 좋은 반면, 포만트 음성합성 방식의 인위적인 결함에 따라 합성된 감정표현 음성 자체가 기계음의 특성을 띠게 된다[2].

결합 합성방식은 미리 구축된 단위음 DB를 합성할 문장에 따라 PSOLA와 같은 음성처리 방법을 통해 결합시킨 후 재생하는 방식을 가리킨다. 이러한 방법을 이용한 감정표현 방식은 합성음이 자연스러운 반면, 감정표현에 사용할 수 있는 파라미터가 매우 제한적이기 때문에 합성음에 표현된 감정의 품질이 떨어지는 단점을 가지고 있다[3].

본 논문은 결합 합성방식을 통하여 Ekman의 기쁨, 슬픔, 화남, 두려움, 혐오, 놀람 등 6가지 감정을 음성 속에 포함시키는 시도를 진행하였다. 먼저 인간의 6가지 기본감정과 일반을 포함한 모두 7가지 감정의 음성 DB를 구축하고 감정에 영향을 미치는 강도, 지속시간,

피치 등 3부류의 특징에 대하여 분석을 진행한 후, 이러한 특징을 일반 음성에 부가함으로써 의도한 감정을 표현하는 방법을 채택하였다. 본 논문에서는 특히 인공생명체의 감정표현을 실현하기 위하여 실제로 인간이 음성을 통해 감정을 표현할 때 피치 궤적이 변화하는 규칙을 분석한 후, 이러한 규칙을 일반 음성에 적용시킴으로써 표현하고자 하는 감정을 구현하였다.

II. 음성 DB 의 구축과 특징 추출

음성 DB 를 구축하기 위하여 20 대 젊은 남성 10 명이 음성 DB 구축과정에 참여하였으며, 음성파일의 녹음은 보통을 포함하여 기쁨, 슬픔, 화남, 두려움, 혐오, 놀람 등 7 가지 감정이 포함되게 하였다. DB 대상문장은 5 개의 부동한 문장을 모두 5 번씩 반복으로 녹음하여 총 1,750 개의 음성 DB 문장을 구성하였다.

음성 DB 의 녹음은 16kHz (sampling rate), 모노채널, 16bit 양자화를 통해 WAVE 파일형식으로 저장하였다. 그리고 저장된 음성파일의 특징을 추출하기 위하여 잡음제거와 지속시간 검출에 불필요한 음성 파일 전후구간 무음부(silence) 제거작업을 진행하였다.

구축된 음성 DB 가 화자의 감정을 얼마나 정확하게 표현하고 있는지를 관찰하고 감정표현 음성처리 모듈을 구현한 후, 성능을 테스트하기 위한 비교수단으로 사용하기 위해 구축된 음성 DB 에 대하여 주관적 평가를 진행하였으며 결과는 표 1 과 같이 나타났다.

표1 주관평가 결과

		인식된 감정					
		기쁨	슬픔	화남	두려움	혐오	놀람
실제 감정	기쁨	69.14	3.01	6.21	4.01	7.82	9.82
	슬픔	2.79	66.07	3.39	9.78	16.17	1.80
	화남	5.40	1.60	66.00	0.80	6.00	20.20
	두려움	5.57	11.73	2.19	58.05	19.48	2.98
	혐오	6.83	16.27	7.03	22.09	42.17	5.62
	놀람	11.02	1.40	14.43	5.41	9.02	58.72
총 인식률							60.1

본 논문에서는 인공생명체의 감정을 표현하기 위하여 음성의 강도(intensity), 지속시간(duration) 및 피치(pitch)에 대한 수정과 음질표현을 위한 주파수 대역 필터링(filtering) 처리를 통하여 구현하였다. 따라서 구축된 음성 DB 에 대한 특징추출 대상항목으로 각각 강도, 지속시간, 피치 궤적(contour) 값을 선택하였으며, 결과는 그림 1~3 과 같이 나타났다[4].

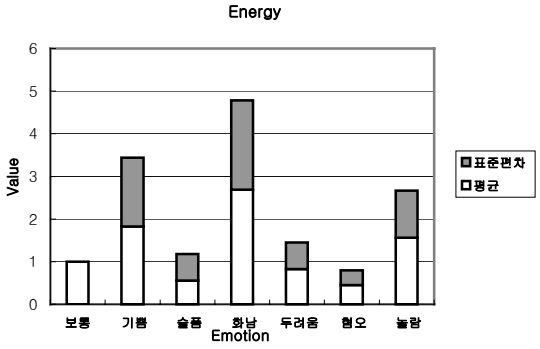

그림1 감정별 에너지 비교

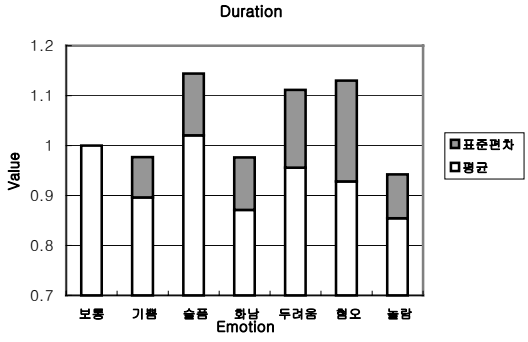

그림2 감정별 지속시간 비교

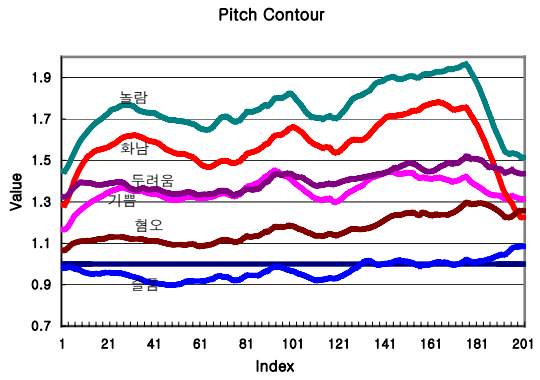

그림3 감정별 피치 궤적의 비교

III. 감정표현 음성처리 모듈의 구현

인공생명체가 음성을 통해 감정을 표현하기 위하여, 자신의 감정상태와 일반 감정이 포함된 음성파일을 입력으로 받고, 음성처리 모듈의 감정변환 과정을 거친 후, 표현하고자 하는 감정이 포함된 음성파일을 출력으

로 내보내게 된다(그림 4).

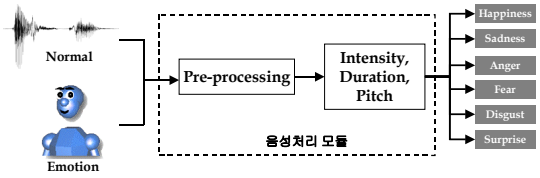

그림 4 감정변환을 위한 음성처리 과정

기쁨의 명쾌한 분위기를 표현하기 위하여 음성의 전처리 과정에서 그림5와 같은 디지털 필터를 통과시킴으로써 저주파수 성분을 억제하였다.

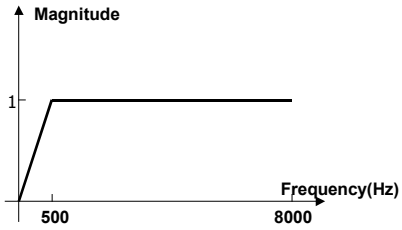

그림 5 기쁨을 표현하기 위한 필터

강도변화는 음성 전체구간에 대해 균일하게 1.5배를 곱하여 구현하였다. 기쁜 감정의 음성은 일반 음성보다 발화속도가 빠른데, 수정계수를 시작부분의 1.0에서 종결부분의 0.9로 변화를 주어 점진적으로 빨라지게 하였다. 피치는 추출된 피치 궤적 값을 간단하게 적용하였다.

슬픔의 울림효과를 강조하기 위하여 일반 음성을 그림6과 같은 필터를 통과시킴으로써, 저주파수 성분을 강화하고 고 주파수 성분을 억제하여 슬픔의 울림소리를 표현하였다.

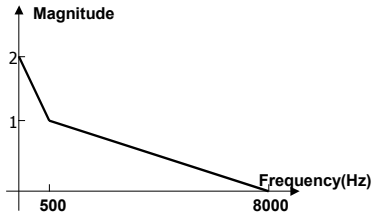

그림 6 슬픔을 표현하기 위한 필터

슬픈 음성의 강도는 일반 음성보다 많이 낮은데, 0.55의 배율을 전체 음성구간에 적용하였다. 지속시간은 수정계수를 1.6로 주어 가장 느린 발화속도를 구현하였다. 피치는 추출된 피치 궤적 값에서 0.2를 낮춰 적용시키는 외에 그림7과 같은 계단 함수(step function)를 적용하여 슬픈 음성에서 나타나는 조음이 불명확한 효과를 추가하였다.

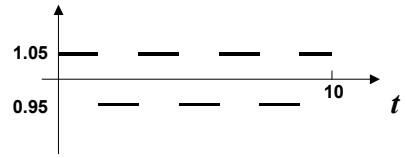

그림 7 슬픔을 표현하기 위한 계단함수

화남을 위한 음성처리에서도 먼저 그림5와 같은 필터를 통해 저주파수 성분을 억제하여 화난 음성에서 목소리가 거세지는 효과를 구현하였다.

화난 음성은 기타 감정이 포함된 음성보다 강도가 훨씬 높으며 전체 음성구간에 2.5의 배율을 균일하게 적용하였다. 특히 화난 음성 속의 무기음을 표현하기 위하여, 0을 평균(mean)으로 하고 자체 음성신호의 100분의 1인 값을 분산(variance)으로 하는 정규분포(Gaussian distribution) 형태의 잡음을 추가하였다. 지속시간은 일반 감정의 음성보다 짧으며 수정계수를 0.9로 선택하였다.

피치 궤적의 특징 추출에서 화난 음성은 놀란 음성과 비교하여 유사한 분포형태를 보였으며, 이들을 구분하기 위하여 앞에서 추출된 피치 궤적 값을 적용한 외에 그림8과 같이 중간부를 강조하는 불연속 선형함수를 적용하여 차별화 시켰다.

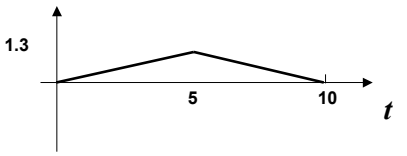

그림 8 화남을 표현하기 위한 불연속 선형함수

두려움을 표현하기 위하여 음성의 강도를 전체 음성구간에 대해 0.6의 배율로 조절하고 지속시간은 0.95의 수정계수를 적용해 발화속도를 높여 주었다.

두려움과 기쁨은 근접한 피치 분포를 나타내고 있으므로 피치 궤적 값을 적용하는데 있어서, 전체 피치 궤적 값을 0.2만큼 증가시켜 차별화 시켰다. 그 외에 그림9와 같은 계단함수를 적용하여 두려움 감정 속의 불규칙적인 발음을 모방하고 불안한 정서를 표현하게 하였다.

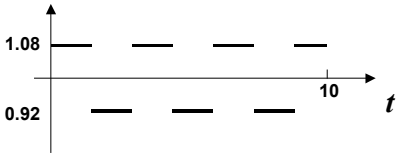

그림 9 두려움을 표현하기 위한 계단함수

혐오를 표현하기 위하여 음성의 강도를 전체 음성

구간에 대해 0.75의 배율로 조절하고, 지속시간은 수정 계수를 시작부분의 1.1에서 종결부분의 1.2로 변화를 주어 끝부분에 말을 끄는 효과를 주었다.

피치의 궤적은 추출된 특징 값을 그대로 사용할 경우, 일반 음성과 구별하기 어려운 점을 감안하여 전체 음성구간에 대하여 피치 궤적 값을 0.3만큼 감소시킨 후 적용하였다.

놀람 감정이 포함된 음성은 끝부분에서 피치를 올리는 대신 강도는 작아지는 경향을 보인다. 이러한 특성을 표현하기 위하여 음성의 강도를 앞에서처럼 전체 음성구간에 대해 단일 값을 적용하지 않고, 그림10과 같이 선형으로 감소하는 변화추세를 추가하였다.

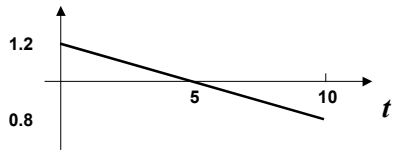

그림 10 놀람을 표현하기 위한 선형함수

지속시간은 전체 감정 중에서 제일 작은 수정 계수인 0.85를 사용하여 가장 빠른 발화속도를 구현하였다.

놀람 감정의 피치 궤적은 먼저 전체 음성구간에 대하여 피치 궤적 값을 0.5만큼 감소시킨 후 적용하였을 뿐만 아니라 그림11과 같은 불연속 선형함수를 적용하여 화난 음성과 명확히 구분 짓도록 하였다.

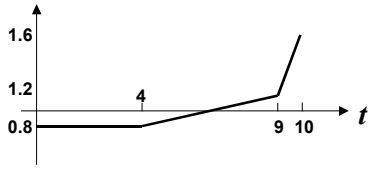

그림 11 놀람을 표현하기 위한 불연속 선형함수

IV. 음성처리 모듈의 테스트

테스트 대상 음성파일은 한국 감성공학 홈페이지 [5]에서 공개한 음성 DB 중, 남성 화자의 일반 감정이 포함된 음성샘플 10 개를 채택하였으며 각 문장은 앞에서 구현한 음성처리 모듈을 통해 기쁨, 슬픔, 화남, 두려움, 혐오, 놀람 등 6 가지 감정이 포함되게 처리하여 모두 60 개의 테스트 대상 음성파일을 구성한 후 음성처리 모듈의 성능검증을 위한 주관적 평가를 진행하였으며 결과는 표 2 와 같이 나타났다.

표 2 감정변환 후 각 감정별 주관평가 결과

		인식된 감정					
		기쁨	슬픔	화남	두려움	혐오	놀람
실제 감정	기쁨	58.0	0.0	11.0	7.0	11.0	13.0
	슬픔	1.0	71.0	2.0	1.0	25.0	0.0
	화남	2.0	0.0	77.0	3.0	1.0	17.0
	두려움	1.0	1.0	0.0	81.0	9.0	8.0
	혐오	21.0	29.0	1.0	6.0	42.0	1.0
	놀람	17.0	0.0	8.0	1.0	13.0	61.0
총 인식률							65.0

V. 결론

본 논문에서는 결합 합성방식을 통한 감정표현의 단점을 극복하기 위하여, 실제로 감정이 포함된 음성 DB를 구축하고 강도, 지속시간, 피치 궤적을 위주로 특징을 추출한 후, 이러한 특징을 일반 음성에 적용함으로써 감정변환이 생성되게 하였다. 그리고 전처리 과정을 통해 포만트 합성방식에서는 쉽게 표현할 수 있었던 명랑함, 울림소리, 조음의 변화 등 특수효과를 모방함으로써 최대한 실제 감정음성에 접근하게 하였다.

구현된 음성처리 모듈을 일반 음성에 적용하여 기쁨, 슬픔, 화남, 두려움, 혐오, 놀람 등 인간의 6 가지 기본감정을 표현하였으며 주관적 평가에서 총 65.0%의 인식결과를 얻었다.

인공생명체가 음성을 통해 감정을 표현하는 방식은 퍼스널로봇, 애완로봇, 음성합성 메일 에이전트 등 다양한 분야에 적용이 가능하며, 본 논문은 이러한 응용에 대비하는 선행연구로서 그 의미를 가진다.

참고문헌

[1] Moulines E., W. Verhelst, “Prosodic Modifications of Speech” in Speech Coding and Synthesis, W. B. Klejin, K. K. Paliwal, eds. 1995, Elsevier, pp. 519-555.

[2] Janet E. Cahn, “Generating Expression in Synthesized Speech,” M.S. thesis, MIT Media Lab, Cambridge, MA, 1990.

[3] Murray I. R., “Emotion in Concatenated Speech,” State of the Art in Speech Synthesis (Ref. No. 2000/058), IEE Seminar on, 2000, pp. 7/1-7/8.

[4] 장국현, “인공생명체의 감정표현을 위한 음성처리,” 한양대학교 석사학위 논문, 2003 년.

[5] http://www.gamsung.or.kr/main.jsp