휴먼-로봇 인터페이스를 위한 TTS의 개발

배 재 현, 오 영 환
한국과학기술원 전자전산학과 전산학전공

Development of TTS for a Human-Robot Interface

Jae-Hyun Bae, Yung-Hwan Oh
Division of Computer Science, Department of EECS, KAIST.
E-mail : {jhbae, yhon@speech.kaist.ac.kr}

Abstract

The communication method between human and robot is one of the important parts for a human-robot interaction. And speech is easy and intuitive communication method for human-being. By using speech as a communication method for robot, we can use robot as familiar way. In this paper, we developed TTS for human-robot interaction. Synthesis algorithms were modified for an efficient utilization of restricted resource in robot. And synthesis database were reconstructed for an efficiency. As a result, we could reduce the computation time with slight degradation of the speech quality.

I. 서론

인간친화형 로봇의 제작에 있어 가장 핵심적인 분야 중 하나는 로봇과 인간 상호간의 의사표현이 인간에게 아주 친숙하고 직관적인 방법으로 수행되어야 한다는 점이다. 로봇이 사용자에게 의사를 표현하는 수단으로써 음성합성은 이러한 조건을 만족시키는 대표적 방법이다. 본 논문에서는 인간 친화형 로봇 내에 탑재될 음성합성 시스템을 개발하였다. 로봇이라는 제한된 자원을 효율적으로 사용하기 위한 방법으로 합성음 생성시 알고리즘을 개선하고[1], 합성코퍼스를 효율적으로 구축하였으며, 로봇내의 각 모듈과 상호 유기적인 작용을 하기 위하여, 각 모듈간의 프로토콜을 일관화 하였다[2]. 개발된 시스템에서는 자연스러운 합성음 생성을 위해 중요한 운율 사건을 중심으로 한 운율처리방식을 간략화 하여 도입하였다[1]. 또한 자주 사용되는 합성단위음 들을 미리 추출해 사용함으로써, 합성음의 음질을 거의 왜곡시키지 않고 합성단위 탐색시간을 줄일 수 있었다[1]. 개발된 시스템의 운율처리 모듈은 CART 기반으로 구현하였으며, 합성방식은 코퍼스 기반 합성방식을 사용하였다[3].

2장에서는 기존 코퍼스 기반 합성방식에 대해 설명하고, 3장에서는 음성코퍼스의 재구축 및 합성알고리즘의 개선에 관하여 기술하고, 로봇내의 각 모듈과의 상호작용을 위한 통신에 관하여 설명한다. 4장에서는 개발된 합성기의 효율성에 관한 실험에 대하여 설명하고, 5장에서 결론을 맺는다.

II. 코퍼스 기반 합성방식

코퍼스 기반 합성방식은 음성에 대하여 인위적인 신호처리 과정을 없애거나 최소화하여 음성 본래의 자연성을 훼손하지 않음으로써, 높은 품질의 합성음을 얻을 수 있는 방법이다. 이 방법에서는 음성에 대한 인위적인 신호처리를 최소한으로 줄이기 위하여 대용량의 코퍼스를 필요로 하게 된다. 대용량 코퍼스를 음소분할하여 데이터베이스로 구축하고, 주어진 텍스트에 해당하는 음소열에 대하여 운율처리 모듈에서 주어진 음소

별 기본주파수, 지속시간 등이 가장 잘 표현된 후보 단위음소들을 조합하여 합성음을 구성한다. 이를 위하여 목적비용과 접합비용이라는 함수가 사용된다. 목적비용이란 주어진 운율에 대해 데이터베이스내의 후보 음소들이 얼마나 일치하는 가에 대한 비용함수이고, 접합비용이란 선택된 후보 음소들을 접합할 때 접합부에서 발생하는 음소 간 특성 불일치를 나타내는 함수이다. 이들 비용의 총 합으로써 합성음을 생성시의 비용함수를 구성할 수 있으며, 아래 식(1)과 같이 나타낼 수 있다.

$$F = \sum_{i=1}^{n} C_i^t + \sum_{j=1}^{n-1} C_{j,j+1}^c \qquad (1)$$

(1)에서 i,j는 합성할 문장의 음소열의 색인이고, 합성할 문장은 n개의 음소로 이루어져 있다고 가정한다. C^t는 목적비용함수(target cost function), C^c는 접합비용함수(concatenation cost function) 을 나타낸다. (1)을 최소화하는 후보 음소열을 추출하여 이를 접합하여 합성음을 구성한다. 목적비용을 계산하기 위해서는 운율생성모듈의 결과인 기본주파수, 지속시간, 에너지 등을 파라미터로 이용하며, 접합비용을 계산하기 위해서는 음향학적 특성의 일치 여부를 알아보기 위해 MFCC, LSF 등을 사용한다. 이때 최적의 후보단위열을 얻기 위하여 코퍼스를 탐색하기 위한 알고리즘으로 Viterbi 탐색 알고리즘이 흔히 사용된다. 이는 그림 1에 나타나 있다.

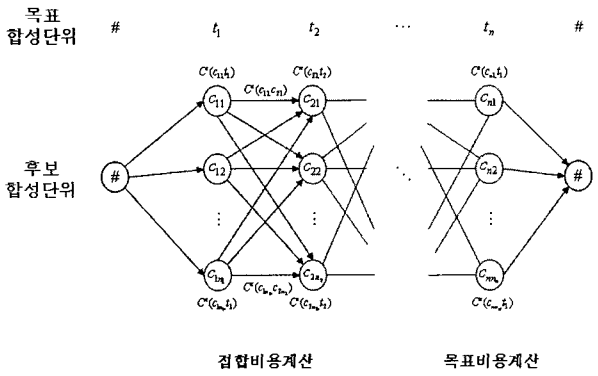

그림 1 후보합성단위 탐색과정

단위음소들의 색인을 구축하기 위하여 전체 코퍼스에 대하여 음소 분할을 수행하여야 하며, 이 작업은 자동 분할 알고리즘을 이용할 수 있으나, 보통 자동분할 후, 수작업으로 보정하게 된다.

단위음소들을 접합하기 위하여 적합한 접합위치를 찾아야 하며, 보통 TD-PSOLA[4] 방식에서 사용되는 피치마커를 기준으로 접합하게 된다. 피치마커를 기준으로 하기 위해서는 전체 음성 코퍼스에 대해 피치를 마킹하여야 하며, 이 또한 자동으로 구축한 후, 수동으로 보정하게 된다.

음성 코퍼스를 구축하기 위하여 녹음 할 문장집합을 결정한다. 이는 보통 원시문장 셋으로부터 PBS(phone balanced sentence)를 추출하여 구성한다[1]. 추출된 녹음 문장집합에 대하여 녹음이 이루어지면 음성 코퍼스를 구축한다. 구축된 코퍼스에 대해 운율학습 및 음소분할, 피치마킹, 에너지 정규화 등을 수행하여 코퍼스를 가공하여 합성에 사용한다. 가공된 코퍼스에 대하여 자주 사용되는 문장, 어절 등을 추출하여 코퍼스의 크기를 줄이고, 계산 시간을 단축한다.

III. 로봇용 음성합성 시스템

본 논문에 사용된 로봇시스템은 펜티엄4를 기반으로 Windows를 사용하는 인터페이스 부와 로봇 몸체의 제어, 보행, 자세제어 등의 모듈이 펜티엄4 기반의 RT-Linux (Realtime Linux)를 사용하는 시스템으로 구성되며, 본 논문은 인터페이스 부에 탑재된다[2]. 인터페이스 부에는 물체인식, 음성인식, 음성합성 몸체 컨트롤부와의 통신 및 제어모듈 등이 탑재되어있다. 제한된 시스템 내에 여러 가지 주요 모듈이 사용되어 품질을 유지하면서 계산 속도를 높이는 과정이 필요하다.

3.1 음성 코퍼스의 개선

본 논문에서는 로봇 내의 제한된 시스템 자원을 고려하여 음성 코퍼스의 크기를 줄이고, 탐색시간을 줄이기 위하여 음소열 선택시 자주 선택되는 음소들을 추출하고, 비슷한 특성을 보이는 음소열 등에 대해 잉여 음소들을 음성 코퍼스 내에서 제거하였다. 또한 거의 선택되지 않는 음소열을 삭제하였다. 이렇게 얻어진 음성코퍼스 2Gbytes를 대상으로 음소별 기본주파수 및 지속시간 분포, 인접음소 등을 고려하여 음소들을 재추출하여 음성코퍼스를 구성하였다. 로봇의 구동환경을 고려하여, 탐색시간을 줄이고, 고품질을 유지하기 위하여 빈번히 사용되는 후보유닛들 중심으로 700Mbytes로 후보단위열을 추출하였다. 추출방법은 그림 2와 같으며 각 트라이폰의 좌, 우 문맥에 따른 접합비용 계산상의 N-best 후보 목록을 만들고, 개체수가 2N개 이상인 트라이폰에 대해 접합비용을 미리 계산하여 목록 작성하여 빈번히 사용되는 후보유닛들을 추출하였다[6].

합성음 트라이폰 문맥 : TP_j -(TP_i)- TP_k

$TP_j \overset{L}{-} TP_i$: 왼쪽 문맥이 TP_j일 때의 N-best TP_i 후보 목록

$TP_i \overset{R}{-} TP_k$: 오른쪽 문맥이 TP_k일 때의 N-best TP_i 후보 목록

$N \le$ 계산시 TP_i 후보수 $\le 2N$

그림 2 빈번히 사용되는 후보유닛의 추출

3.2 합성알고리즘의 개선

운율생성 모듈로부터 넘겨받은 운율 정보에 적합한 후보단위열을 탐색할 때, 운율이벤트의 발생 유무를 중심으로 후보 선택시 가중치를 부여하는 방법에 착안하였다[1]. 이 방법을 간략화 하여 운율 궤적에 대하여 그림 3에서와 같이 운율 궤적의 국부최대화 지점을 선택하여 이를 운율이벤트가 발생한 지점으로 간주하였다. 음성 코퍼스로부터 후보 음소열을 탐색할 때에는 (2)와 같이 음소별로 가중치를 두어 음소열을 선택하였다. 운율 이벤트 간에는 자연스럽게 이어주는 방법으로 가중치를 낮추어 후보단위열을 선택하였다. 목표비용 가중치를 높이고, 이들 지점 간에는 자연스러운 연결이 되는 접합비용에 가중을 두어 후보열을 선택하였다.

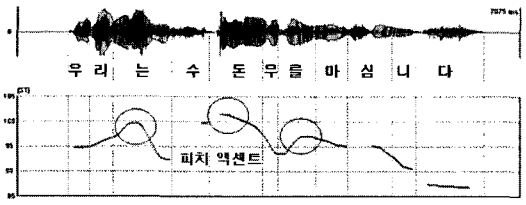

그림 3 운율 이벤트의 선택

그림 2에서 w_k는 국부 최대화 지점에서 1이 되며, 국부 최대화 지점 간 가운데 지점에서 0이 되며, 그 사이에서는 0과 1 사이의 값을 가진다. 이는 (2)와 같다.

$$F = w_k C^t_k + (1 - w_k) C^c_{k-1,k} \quad (2)$$

(3)에서 $D_L(k)$는 k번째 음소에서 왼쪽으로 제일 가까운 이벤트 지점의 음소 위치이며, $D_R(k)$는 오른쪽으로 제일 가까운 이벤트 지점의 음소 위치를 나타낸다. (2)와 같이 이벤트 지점 사이를 목적 비용에 대해 가중치를 부여하며, 이에 대한 보상 개념으로 접합비용에 대한 가중치를 부여한다.

$$w_k = \begin{cases} 1, \text{if } D_R(k) = 0 \\ 0, \text{if } D_L(k) = D_R(k) = m \neq 0 \\ \dfrac{k - D_L(k)}{m - D_L(k)}, \text{if } D_L(k) < D_R(k) \\ \dfrac{D_R(k) - k}{D_R(k) - m}, \text{if } D_L(k) > D_R(k) \end{cases} \quad (3)$$

3.3 신호의 접합 및 신호합성

기존 코퍼스 기반 합성방식에서는 수동으로 조절된 피치마커를 기준으로 왼쪽 음소와 오른쪽 음소를 접합하였으나, 본 논문에서는 이를 자동화하여 좌,우 음소의 접합부분의 유사도를 비교하여 최적의 접합점을 자동으로 구하였다. 비교시의 구간은 왼쪽음소의 마지막 2배 기본주기 구간과 오른쪽 음소의 처음 2배 기본주기 구간으로 설정하였다. 유사도는 음성신호의 상관계수로 하였다. 그 후, 속도를 개선하기 위하여 좌, 우 신호의 차이가 최소가 되는 지점으로 선정하였다.

본 논문에서 구현한 코퍼스 기반 신호합성부의 구조는 그림 4와 같다. 대규모음성 코퍼스에서 적절한 음소들을 추출하여 실제 신호 합성시 사용되는 음성 코퍼스를 구축하고, 운율정보로부터 식(2)를 이용하여 후보 음소열을 추출하고, 이를 접합하여 합성음을 생성한다.

음성 코퍼스를 위해 약 6000문장이 사용되었으며, 전문 성우의 음성을 16kHz, 16bit 양자화 하여 사용하였다. 합성음의 청취테스트를 위하여 피실험자 5명을 대상으로 10개의 문장을 들려주고 1 ~ 5점 사이의 점수를 선택하도록 하였다. 테스트에 쓰인 문장은 다음과 같다.

- 재미있는 이야기를 해 봅시다.
- 그중에 야심만만한 젊은 개구리 한마리가 늘 큰소리를 쳐댔다.
- 그러나, 개구리들은 한번도 웅덩이를 벗어나 보디 않았기 때문에, 떠나가는 것이 불안하게 느껴졌다.
- 하루종일 사람들이 차를 타고 다니기 때문에 몹시 위험하다.
- 대전시의 교통상황을 말씀드리겠습니다.
- 충대 오거리 부근의 교통이 혼잡합니다.
- 학교에 있는 꽃밭에 많은 꽃들이 피어있습니다.
- 그는 후회를 했지만, 이미 아무런 방법도 찾아낼 수가 없었다.
- 여름의 도로한복판은 구운 돌처럼 뜨거웠고, 수많은 차들이 내뿜는 매연으로 열섬현상이 발생하였다.
- 언젠가는 저 아래에 있는 큰 연못으로 가서 살고 싶었다.

테스트 결과 700Mbytes 코퍼스로 구성된 합성기의 경우 MOS 3.23점을 얻었으며, 기존 2Gbytes 코퍼스로 구성된 합성기의 경우 3.32점을 획득하였다. 합성 속도

는 기존 2Gbytes 합성기 대비 약2배정도 빨라진 것을 확인할 수 있었다. 이는 합성단위 선선택 및 음성 코퍼스의 재구축으로 인한 효과로 분석된다. 또한 기존 피치마커를 이용한 합성방식에 비해 유사도 측정을 통한 접합점 검출방식을 사용함으로써 이로인한 속도감소 효과가 있는 것으로 나타났다.

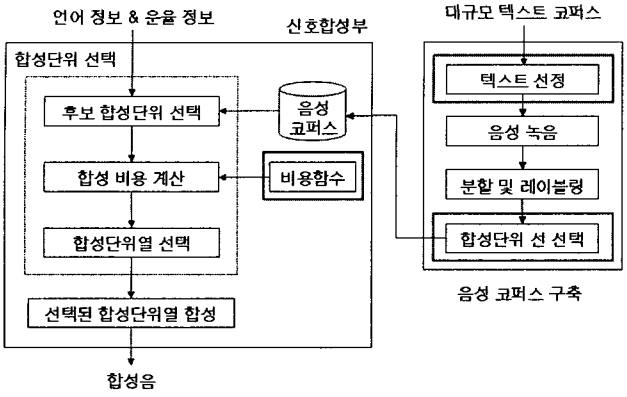

그림 4 구현된 신호처리 및 음성코퍼스의 구성

Ⅳ. 로봇내 모듈간 상호작용

음성합성 시스템이 속한 인터페이스 부에는 음성인식, 합성, 물체인식 등과 같은 모듈이 속해있으며 각 모듈은 Central Planner를 통해 서로 통신한다. 그림 5는 그 중 합성시스템 모듈을 나타낸 것으로, Central Planner로부터 합성할 문장을 넘겨받아 로봇에 내장된 스피커로 출력한다.

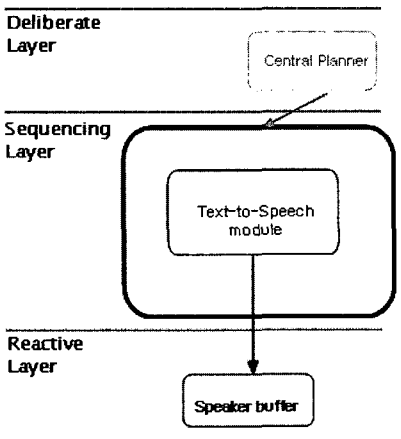

그림 5 인터페이스 모듈내의 합성시스템

Central Planner와의 통신은 TCP/IP를 통하여 구현하였으며, Embot ID 3으로 동작한다.

Ⅴ. 결론

본 논문에서 제안한 음성 합성기술은 기존의 소용량 합성기술에 비해 보다 자연스럽고, 풍부한 음성을 생성할 수 있으며, 로봇이라는 제약된 컴퓨팅 환경에서 여러 가지 무거운 모듈이 동작하는 환경에서 원활하게 작동할 수 있도록 설계되었다. 소형 전자기기나 가전 등과 같은 Embedded 분야에 적용하기 위해 코퍼스 기반 방식이 아닌 수메가 이내의 소용량 기반방식으로 코퍼스 기반 방식과 준하는 합성음을 생성하는 연구를 수행중이다.

감사의 글

본 연구는 과학기술부의 지원을 받아 2006년도 국가지정연구실을 통해 수행되었음

참고문헌

[1] Heo-JIn Byeon and Yung-Hwan Oh, "An Event-driven f0 Weighting for Prosody Control in A Large Corpus-based TTS System", IEEE Signal Processing Letters, Vol. 11, No. 2, pp. 262-265, February 2004.

[2] 김종환, 오영환 외, "네트워크 환경에서의 유비쿼터스 로봇의 구현", J. Control, Automation and System Engineering Vol. 11 No. 8, August 2005

[3] Sangho Lee and Yung-Hwan Oh, "Tree-based modeling of intonation", Computer Speech and Language, Vol. 15, No. 1, pp. 75-98, January 2001.

[4] 오영환, "음성합성 기술 개발 현황",제 11회 음성통신 및 신호처리 워크샵 논문집, pp. 271-274. 1994

[5] A. J. Hunt and A. W. Black, ""Unit selection in a concatenative speech synthesis using a large speech database,"" in *Proc. ICASSP*, Vol. 1, pp.1:373--1:376. 1996,

[6] 변효진, "대규모 코퍼스 기반 TTS 시스템에서의 운율제어", Phd Thesis, KAIST. 2004