시청각 기반 HRI 컴포넌트 상용화 서비스 현장 성능 평가 및 환경분석

지수영 | 김혜진 | 김도형 | 윤호섭
한국전자통신연구원

요 약

본고에서는 지능형 서비스 로봇의 상용화 단계에서 가장 현실적으로 적용 가능한 대표적인 HRI기술(얼굴검출, 화자 성별구별, 음원추적)에 대하여 상용화 서비스 현장에서의 성능평가 결과를 제공하고, 현장을 분석하여 사용자에게 가이드라인을 제공함과 동시에 최적의 상용화 서비스 제공을 위한 사용자와 로봇간 HRI 기준 및, 공공로봇 플랫폼 적용을 통한 로봇 서비스의 Needs 파악과 상품기획력의 극대화를 목적으로 성능평가에 따른 환경분석을 제안한다.

I. 서 론

로봇산업의 진화과정을 보면 1980년대 자동차나 전자 산업 등의 노동집약적 산업의 발달에 따라 로봇의 생산현장 투입을 통해 산업용 로봇 시장 성장이 이루어졌고 1990년대 산업용 로봇 시장성숙기에 따른 응용전환기를 거쳐 2000년 이후 생활환경의 변화와 고령화 사회의 진전과 특히 network기반의 IT기술의 도약적인 발전에 따른 지능형 서비스 로봇의 시장이 열림으로 신규 서비스 로봇 시장의 확대가 가능 함에 따라 우리의 생활 속으로 이제 로봇이 들어와 인간과 함께 공존하는 새로운 로봇 혁명기를 맞고 있다.

이러한 로봇의 출현은 인간과 환경의 연결고리인 네트워크기술이 "인간중심의 네트워크 체계"인 Ubiquitous Network 사회의 도래를 통하여 생활 속에서 사람과 네트워크와의 새로운 관계를 가져옴으로 가능해졌고 개별 기능 위주에서 기능의 네트워킹화를 통하여 Agent Service화로 진화하면서 Digital Convergence 및 Networking과 로봇의 서비스 형태가 "Anytime, Anywhere"인 지능형 Agent Service로 진화가 진행되기에 기술과 인간욕구가 결합된 본격적인 인간중심 기술의 구현인 신 성장동력으로서의 Network기반의 지능형 서비스로봇의 실현이 가능하다.

이러한 사회적인 기술의 환경변화에 따라 언제 어디서나 나와 함께 하면서 나에게 필요한 서비스를 제공하는 로봇산업화를 위한 대안적 모델인 URC(Ubiquitous Robot Companion) 개념의 로봇인 Network Based Intelligent Service Robot이 나오게 되었고[1,2] 이는 단 품 로봇의 제약성을 극복하고 다양한 고도의 기능과 서비스의 제공이 가능하며 Mobility와 Human Interaction이 향상된 로봇 시스템으로 진화하였다.

URC 개념의 로봇은 기존의 로봇개념에 Network의 부가기능을 추가하여 Usability 향상과 Killer Application 발굴과 서비스 기능을 향상시킴으로 소비자의 Benefit을 확대하였고 Mobile Platform을 갖춘 Multi Function로봇으로의 개념확대와 플랫폼 가격의 인하와 다양한 기술적 접근을 통한 Robot Cost 문제를 해결 할 수 있다.

로봇을 단순한 기계적인 기능으로 보는 전통적 시각에서 문화적으로 메시지의 전달이 가능하여 인간과 로봇의 상호작용이 가능한 새로운 미디어로서의 단말 기능이 가능한 동반자 로봇이라는 문화적 상식이 통할 때 현재의 로봇 장벽들에 대한 새로운 인식을 통한 시장 창출이 가능하리라 보고 우리의 삶 깊이 동반자로서의 로봇이 이제 자연스럽게

다가오는 시대가 곧 열리리라고 기대한다.

1.1 배경

본고에서는 단순 서비스가 아닌 지능화된 로봇 서비스의 요구가 증대하고, 능동형, 맞춤형, 지능형 서비스를 위한 HRI 기술의 발전이 빠르게 진행되고 있으며, HRI 기술에 대한 현장에서의 실제 성능평가 요구가 점점 커지고 있고 최적의 서비스 제공을 위한 사용자와 로봇간 HRI 기준 개발이 시장에서 요구됨으로 인하여, 공공로봇 플랫폼 적용을 통한 로봇 서비스에 대한 Needs 파악 및 상품기획력 극대화와 현재 상용화 되거나 시범사업에 적용되고 있는 지능형 로봇은 인간을 인지하여 인간과 자연스럽게 교감하는 HRI(Human Robot interaction)기술의 수요에 의하여 로봇이 한 단계 발전된 고차원의 지능화 된 서비스를 제공하기 위해서는 HRI 기술이 필수적으로 요구된다.

하지만 실제 로봇이 상용화 서비스를 하고 있는 현장에서의 HRI S/W의 성능에 대한 실험결과는 없어, 로봇개발 업체 및 실사용자에게 HRI S/W의 성능에 대한 적절한 정보를 제공하지 못하고 있다.

따라서, 본고에서는 상용화 단계에 가장 접근한 대표적인 시청각 기반 HRI 기술(얼굴검출, 화자성별구별, 음원추적)[3-10]에 대하여 현장에서의 성능 평가 결과를 제공하고, 환경을 분석하여 사용자에게 가이드라인을 제공함을 그 목적으로 한다.

1.2 상용화 서비스 HRI 성능평가 항목 분류

본고에서는 시청각 기반 HRI 상용화 서비스의 성능평가를 위하여 아래와 같이 5가지의 분류로 집중하여 성능평가를 수행함.

1) 얼굴검출, 화자성별구별, 음원추적 성능평가 모델 개발
2) 음원추적용 성능평가 시스템 구축 및 현장 성능평가
3) 얼굴검출, 화자성별구별 현장성능평가용 데이터 베이스 구축
4) 얼굴검출, 화자성별 구별 성능평가
5) 성능평가에 따른 환경분석

II. HRI 컴포넌트 성능평가 방법론

본고에서는 로봇이 한 단계 발전된 고차원의 지능화된 서비스를 제공하기 위해서는 HRI기술이 필수적으로 요구되는 바 현재 상용화 되거나 국민로봇 시범사업에 적용되고 있는 전략적인 HRI 기술(얼굴검출, 화자성별 구별, 음원추적) 에 대하여 성능평가 방법론을 제시한다.

2.1 상용화 서비스 현장 데이터 수집

2.1.1 얼굴 검출 컴포넌트 성능평가

국민로봇 상용화 서비스 현장의 데이터를 수집하기 위하여 영상을 획득할 수 있는 카메라가 탑재된 로봇에서 영상을 획득하여 데이터베이스를 구축한다. 획득 받은 영상은 프레임 단위로 저장되고 구축된 데이터베이스는 획득 받은 영상과 얼굴이 있는 위치가 기록된 그라운드트루스 정보로 구성한다

1) 데이터베이스 구축

서비스 현장에 대한 얼굴검출 컴포넌트 데이터베이스 구축은 국민로봇 2차 시범사업 사이트인 총 3~5 사이트에서 구동되는 로봇을 통하여 영상을 획득하고 영상 획득 S/W를 로봇 또는 원격지의 로봇영상 획득 서버에 탑재하여 구동한다. 영상 획득 S/W는 각 사이트에서 총 10일간 구동하고 시간에 따른 조명의 변화를 영상에 반영하기 위하여 하루에 3회, 1시간씩 자동 구동한다. 영상 획득 속도는 1frame/sec이고 획득되는 영상은 640x480 컬러 영상으로 입력된다. 10일 후 한 사이트에서 획득되는 영상은 총 108,000 프레임으로 예상된다. GroundTruth는 얼굴의 있는 위치를 text 파일로 저장하고 얼굴의 위치는 얼굴의 가운데 지점을 수동으로 마킹한 x,y 좌표로 표현한다.1프레임당 1개의 그라운드트루스가 존재.

2) 평가방법

현장에서 구축된 데이터베이스를 기반으로 표1과 같은 항목으로 얼굴검출 성공률을 측정한다. 총 테스트 프레임은 최대 5 사이트로 각 사이트 당 108,000 프레임으로 총 540,000 프레임으로 평가한다

〈표 1〉 얼굴검출 컴포넌트 성능평가표

촬영 장소	촬영 일자	촬영 시간대	총 얼굴 개수	정검출 개수	얼굴누락 개수	오검출 개수	정검출률
A	1일	10~11	100	90	10	3	90%

2.1.2 화자성별 구별 컴포넌트 성능평가

상용화 서비스를 위한 화자성별 구별 컴포넌트의 성능평가는 음성을 획득할 수 있는 마이크가 탑재된 로봇에서 음성을 획득하여 데이터베이스를 구축하고 획득받은 음성은 문장 단위로 저장된다. 구축된 데이터베이스는 획득 받은 음성과 발화자의 성별이 기록된 그라운드트루스 정보로 구성한다. 구축된 데이터베이스에 대하여 화자성별구별 컴포넌트의 성능을 정량적으로 평가한다.

1) 데이터베이스 구축

서비스 현장에 대한 화자성별 구별 컴포넌트 데이터베이스 구축은 음성획득 S/W는 영상획득 S/W와 같이 구동되므로, 영상획득 S/W가 구동되는 장소, 시간, 조건 등이 동일하다. 2초 이상 발화되는 음성을 문장단위의 raw 파일로 저장한다. GroundTruth 해당 raw 파일의 성별을 text 파일로 저장한다. 1 문장 당 1개의 그라운드트루스가 존재한다.

2) 평가방법

구축된 데이터베이스로 표2의 화자성별 구별 컴포넌트 성능평가표 항목에 따라 화자성별구별 성공률을 측정한다.

〈표 2〉 화자성별 구별 컴포넌트 성능평가표

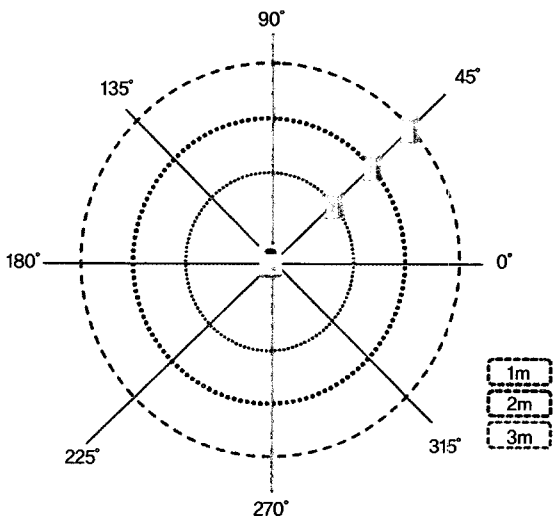
녹음 장소	녹음 일자	녹음 시간대	남자		여자		성별구별 성공률
			총문장개수	정구별개수	총문장개수	정구별개수	
A	1일	10~11	40	30	60	50	80%

2.1.3 음원추적 컴포넌트 성능평가

음원추적 컴포넌트의 현장 성능평가방법은

해당 사이트의 평균 소음 정도를 측정하고 표기하여, 평가결과와 함께 제공한다. 하나의 대표 block에서 스피커와 로봇과의 상대적 위치를 아래 그림과 같이 8방향, 3거리로 구성하여, 평가한다. (24set으로 구성)

스피커와 로봇은 지정된 곳에 미리 위치하며, operator가 버튼 등을 클릭하면 음원추적 시스템이 동작을 시작함과 동시에 스피커가 음성을 발화한다. 시스템이 추적결과를 출력하면, 인식결과를 기록한다.



(그림 1) 음원추적 성능평가를 위한 스피커와 로봇간의 위치

얼굴검출 성능평가방법은 각 현장의 로봇에 영상획득용 S/W 장착하고 총 10일간 영상 DB를 구축한다. (총 236,206장) 구축된 DB를 회수하여 그라운드트루스 생성 (눈좌표 수동 마킹)한 후 최종 구축된 DB를 이용 얼굴검출 성능 평가를 시행한다.

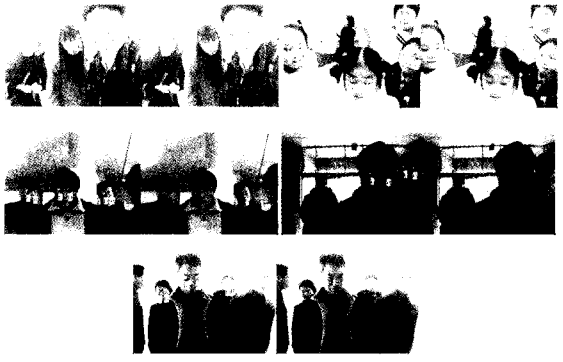
〈표 3〉 얼굴검출 성능평가 결과

평가현장	총 영상수	총 얼굴수	정검출률 (%)	검출실패율 (%,FR)	오검출률 (%,FA)
누리마루(부산)	55,884	14,004	97.44	2.56	0.65
시립미술관(대전)	15,983	9,383	98.41	1.59	0.94
레스토랑VIPS(서울)	95,950	11,666	90.95	9.05	0.11
KT텔레캅 본사(서울)	68,389	9,699	97.48	2.52	2.17
합 계	236,206	44,752	96.07	3.93	0.97

-정 검출률(%) = 100 x 정검출수/총얼굴수

-검출실패율(%) = 100 x 검출실패수/총얼굴수

-오검출률 = 100 x 오검출수 / 총영상



(그림 2) 얼굴검출 성능평가 결과영상

2.3 화자성별구별 성능평가 분석

화자성별구별 평가 방법은 각 현장의 로봇에서 음성 인식 시에 녹음된 음성데이터를 입수한 후 각 음성데이터를 문장 단위로 분리하여 음성 DB를 구축한다. (총 361 문장) 최종 구축된 DB를 이용 〈표 4〉와 같이 화자성별 구별 성능 평가를 시행한다.

〈표 4〉 화자성별구별 성능평가 결과

평가 현장	총 문장개수	구별 성공률(%)
누리마루(부산)	102	93.98
시립미술관(대전)	95	96.10
레스토랑VIPS(서울)	164	77.552
합 계	361	89.212

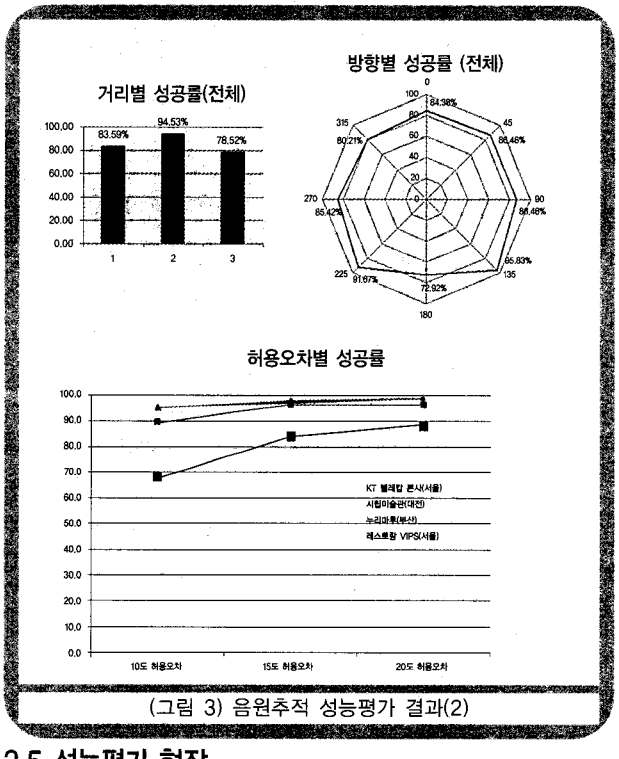
2.4 음원추적 성능평가 분석

음원추적 성능평가는 ETRI에서 구축한 음원추적 시스템으로 현장 평가한다. 로봇 마이크와 스피커와의 거리는 1m, 2m, 3m로 구분하여 진행된다. 시험데이터로는 무소음 환경에서 사전에 녹음을 하고 화자의 구성은, 총 3명 (성인남자 1, 성인여자 1, 아동 1)으로 정하여 다음과 같은 평가항목들로 수행한다.

각도는 로봇을 정면으로 바라보도록 설치하고 박수소리는 1개, 문장은 3 문장으로 화자별로 한 번씩 녹음하여 수행한다. 재생 신호 크기는 녹음된 문장을 재생하였을 때 신호의 크기를 1m 떨어진 곳에서 소음계로 측정시 70~80dB이 나오는 크기로 설정하여 수행한다. 스피커와 로봇과의 상대적 위치를 (그림 3)과 같이 8방향, 3거리로 구성하여, 평가 한다. (24set으로 구성) 음원추적 결과가 ±10도 내이면 추적 성공으로 간주한다.

〈표 5〉 음원추적 성능평가 결과(1)

평가현장	추적 성공률 (%)	평균 소음 (dB)	전체오차 평균 (degree)	성공샘플 오차평균 (degree)	실패샘플 오차평균 (degree)
누리마루(부산)	95.31	61.1	4.46	3.33	27.41
시립미술관(대전)	89.58	59.0	8.00	3.70	46.10
레스토랑 VIPS(서울)	67.71	61.0	16.00	3.97	41.62
KT 텔레캅 본사(서울)	89.58	56.5	6.00	3.43	29.60
합 계	85.55	59.4	8.61	3.50	39.09



(그림 3) 음원추적 성능평가 결과(2)

2.5 성능평가 현장

시청각 기반 HRI 컴포넌트 현장 성능평가는 국민로봇 2차 시범사업 현장 중, 〈표 6〉과 같이 4개의 현장에서 현장 성능평가를 수행함.

〈표 6〉 상용화 시범사업 성능평가 현장

컨소시엄	평 가 현 장	로봇 제공업체
KT	누리마루,부산	로보테크
	시립미술관, 대전	다사로봇
ED	레스토랑 VIPS (서울 어린이대공원점)	ED로봇
DU	KT 텔레캅 본사 (서울 구로)	DU로봇

성능평가 현장의 실제모습은 (그림 4)에 현장에 투입된 성

능평가 플랫폼과 다양한 로봇은 (그림 5)와 같다.

(그림 4) 성능평가 현장

(그림 5) 성능평가 플랫폼 및 장비

III. 결 론

로봇이 한 단계 발전된 고차원의 지능화된 서비스를 제공하기 위해서는 HRI 기술이 필수적으로 요구된다. 현재 상용화되거나 시범사업에 적용되고 있는 지능형 서비스로봇은 인간을 인지하여 인간과 자연스럽게 교감하는 HRI기술의 탑재가 저조한 형편이다. 실제 로봇이 상용화 서비스를 하고 있는 현장에서의 HRI S/W의 성능에 대한 실험결과는 없어, 로봇개발 업체 및 실사용자에게 HRI S/W의 성능에 대한 적절한 정보를 제공하지 못하고 있다. 따라서 본고에서는 지능형 서비스 로봇의 상용화 단계에서 가장 현실적으로 적용 가능한 인간로봇상호작용(HRI)기술인 얼굴검출, 화자성별구별, 음원 추적에 대하여 상용화 서비스 현장에서의 성능평가 결과를 제공하고, 현장을 분석하여 사용자에게 최적의 가이드라인을 제공함과 동시에 상용화 서비스 제공을 위한 사용자와 로봇간 HRI 기준을 개발하고, 공공로봇 플랫폼 적용을 통한 로봇 서비스의 Needs 파악 및 상품 기획력의 극대화를 목적으로 성능평가에 따른 환경분석을 제안하였다.

실제환경에서의 얼굴검출 컴포넌트 성능, 화자성병구별 컴포넌트 성능 그리고 음원추적 컴포넌트 성능평가 결과 조명 잡음이 열악한 환경에도 불구하고 높은 인식성능을 보였고, 이를 바탕으로 HRI 컴포넌트 기술은 실제 환경에서 적용이 가능하며 상용화가 가능한 수준임을 알 수 있었다. 다만 (그림 6)과 같이 극한 조명 및 소음환경에서는 성능개선이 필요함을 알 수 있었다. 개선을 위한 방안으로 조명 처리 및 잡음제거 S/W 모듈 개선과H/W 장착을 통한 성능 개선, 다채널 음성 보드 및 마이크의 장착 및 역광보정 카메라의 부분 적용이 필요하다 하겠다.

(그림 6) 극한환경에서의 실패 데이터

참 고 문 헌

[1] 조영조, " 지능형 서비스 로봇과 URC", ETRI 주간기술동향, No. 1150, pp. 29-38, 2004.

[2] 김현" URC에서의 소프트웨어 로봇 기술", 한국통신학회지, V.21 No.10, pp. 36-43, 2004.

[3] 김도형, "다중 특징 결합과 유사도 공간을 이용한 SVM 기반 얼굴 검출 시스템", 한국정보과학회논문지, V.31 No.6, 2004.6.

[4] DoHyung Kim, "A Non-Cooperative User Authentication System in Robot Environment", IEEE Transactions on Consumer Electronics, V.53 No.2, 2007.

[5] Ho Sub Yoon, "A robust head detection method for human tracking", IROS 2006, 2006.

[6] 곽근창,지수영, "소프트웨어 로봇을 위한 인간로봇 상호작용", 대한전자공학회지, 2006.

[7] DoHyung Kim, "Face Identification for Human Robot Interaction:Intelligent Security System for Multi-user Working Environment on PC", ROMAN 2006, 2006.

[8] KyungSuk Bae, "Speaker' s gender identification for human-robot interaction", SIGMAP06, 2006.

[9] Mikyoung Ji, "Text-Independent Speaker Recognition for Ubiquitous Robot Companion", URAI2006, 2006.

[10] Sungtak Kim, "Noise-Robust Speaker Recognition Using Subband Likelihoods and Reliable-Feature Selection", ETRI Journal, V.30 No.1, pp. 89-100, 2008.2.

지 수 영

1986년 충북대학교 전산학학사
1988년 충북대학교 공학석사
2005년 고려대학교 이학박사
2006년 ~ 2007년 USC 파견연구원
1991년 ~ 현재 한국전자통신연구원 융합기술미래기술연구부 책임연구원
관심분야 : 인간로봇상호작용(HRI), 지능형 서비스 로봇, 네트워크 서비스 로봇, OMG Robotics 국제표준화

김 혜 진

2001년 포항공과대학교 공학사
2003년 포항공과대학교 공학석사
2004년 ~ 현재 한국전자통신연구원 연구원
관심분야 : 화자인식, 머신러닝

김 도 형

2000년 부산대학교 이학사
2002년 부산대학교 이학석사
2002년 ~ 현재 한국전자통신연구원 연구원
관심분야 : 얼굴인식, 로봇비젼, 영상처리

윤 호 섭

1989년 숭실대학교 공학사
1991년 숭실대학교 공학석사
2003년 KAIST 공학박사
1991년 ~ 1998년 KIST 시스템공학센터 선임연구원
1998년 ~ 현재 한국전자통신연구원 지능형로봇연구단 책임연구원
관심분야 : 로봇비전, 영상처리, 음성처리, 패턴인식