시청각 정보를 활용한 음성 오인식률 개선 알고리즘

Improvement Algorithm for Audio Recognition Error Rate Utilizing Audio-Visual Information

*#이광희[1], 고우현[1], 지상훈[1], 남경태[1], 이상무[1]
*#K.H.Lee(leekh@kitech.re.kr)[1], W.H.Ko[1], S.H.JI[1], K.T.Nam[1], S.M.Lee[1]
[1] 한국생산기술연구원 지능형로봇연구부

Key words : Auditory-Visual Recognitioin, Intergration, Medical Robot, Lip Detection, HRI

1. 서론

정교한 제어기술을 이용한 산업용 로봇은 인간에게 물질적 풍요를 가져다 주었으며 급격히 발전해 가는 로봇기술은 인간의 삶 전반에 보급되어 지고 있다. 또한 오늘날 노인 인구의 증가와 함께 의료용 로봇에 대한 관심과 개발에 대한 연구가 많이 진행되고 있다. 의료용 로봇은 사람에 비하여 정밀하고, 안정되게 제어될 수 있으며, 작업에 대한 반복성이 우수하여 수술실에서 집도의의 명령을 따라 수술을 보조하거나, 영상 가이드를 해주는 수술 보조 로봇 등이 개발되고 있다.[1]

로봇은 의사의 지시를 조이스틱, 터치스크린, 버튼등과 같은 다양한 인터페이스를 통해 명령을 받을 수 있다. 하지만 의사의 손에 다양한 수술용 도구를 들고 있다면 의료용 로봇을 동작시키기 위해 수술 도구를 내려 놓을 수 상황이 생기거나 보조의에게 요청을 할 수 밖에 없을 것이다.

이러한 문제를 해결하기 위해 음성정보나 비전정보를 이용하여 사람과의 상호작용을 통해 보다 다양한 정보들을 얻어 낼 수 있다. 예를 들어 의사의 손과 발을 사용하지 않고 로봇과 자연스러운 의사소통 방법으로 음성 정보를 이용하거나 비전정보를 의사의 머리 부착되어 있는 패치의 움직임을 인식하여 수술을 도와주는 방법이 연구가 되어지고 있다.[2]

비전을 이용하면 영상처리를 위한 연산이 많아 빠른 명령 입력이 어렵고 수술자의 움직임의 실수로 로봇의 오작동으로 큰 문제가 생길 수 있다. 이에 반해 음성 정보는 상대적으로 적은 크기의 음성 신호를 처리하여 다양한 명령어를 인식할 수 있다.[2]

하지만 음성 신호의 경우 실제 동작 환경 내 발생하는 음성 노이즈에 의해 데이터가 변질되어 성능이 저하 될 뿐만 아니라 오인식이 발생할 확률이 높아진다.

노이즈에 의한 오인식을 보완하기 위하여 시청각 인식 방법을 이용하여 획득한 영상에서 실시간적으로 입술의 움직임을 추출하여 입술이 움직일 경우 음성인식기의 결과값을 사용하는 등 다양한 잡음에 강인한 성능을 갖는 인식 방법들이 연구가 진행되어 지고 있다.[2,3,4]

인간의 생명을 다루는 의료용 장비에서 가장 중요한 것은 정확도이다. 데이터가 손실되어 없어진다면 문제가 생기지 않겠지만 오인식 되어 원하지 않는 동작을 한다면 큰 문제가 될 것이다.

본 논문에서는 음성 노이즈에 강인한 여러 방법들 중 음성정보와 시각정보를 함께 사용하는 시청각 인식 방법을 이용해 오인식률을 낮추는 방법을 제시하고자 한다.

비전정보에서 모음을 정확하게 추출하기 위하여, 입술과 입술 안쪽영역의 특징을 분석하였고, 사전에 모음 각각에 대한 가중치와 임계치를 결정하는 문제를 다룬다. 또한 음성정보에서 추출한 모음과 비전 정보에서 얻은 모음의 가중치를 비교하여 임계치와 비교하여 오인식률을 낮출 수 있다.

2장에서는 비전시스템에서 획득한 영상으로 부터 모음을 추출하는 방법을 설명하고, 3장에서는 영상정보에서 추출한 모음과 음성정보에서 추출해 낸 모음을 비교하여 오인식을 검출하는 방법을 소개한다. 4장에서는 제안한 방법을 이용하여 실험한 결과를 보여주며, 5장에서는 결론을 내린다.

2. 입술영역 추출

한글은 자음과 모음이 결합하여 음절을 만들어내며 사람은 적절한 혀의 위치와 입모양을 만들어 발음하게 된다. 자음발음은 주로 혀의 변화와 위치에 영향을 받고, 모음은 입모양에 영향을 받는다. 특히 단모음 (ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ, ㅐ, ㅒ)은 자음과 구별이 잘되는 입모양을 가지고 있는 특징으로 비전 정보를 이용하여 추출 할 수 있다. 단모음을 추출하기 위한 영상처리는 Fig 1과 같이 크게 3단계로 이루어진다.

Fig. 1 Procedure of finding features of the lip segment

첫 번째 단계, 비전 데이터에서 입 영역은 사람 얼굴 하단 1/3 영역에 존재한다는 것을 이용하여 얼굴 위치를 찾아낸 후 입술을 ROI로 추출한다.

두 번째 단계, 상으로부터 추출하고자 하는 색상을 고려하여, 색상과 보색의 관계를 가지는 RGB 3채널 중 녹색을 가산 연산해 줌으로써 입술을 선명하게 해주었다. 또한 히스토그램을 이용하여 영상의 밝기를 보정하고 입술영역을 뚜렷하게 하기 위해 픽셀 값 분포를 정규화(Equalization) 하였다.

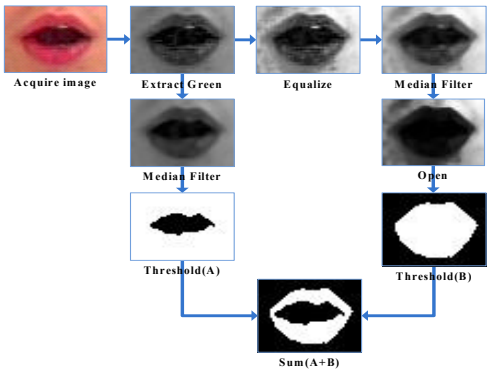

Fig. 2 Pre-processing procedure for the detection of the Lip area

입술 주위의 잡음을 제거하기 위해, 통상적으로 임펄스 잡음을 없애거나 이미지의 윤곽선을 뚜렷하게 처리를 수행하기 이전에 많이 사용하는 미디언 필터를 사용하였고, 잡음이 감쇠된 것을 확인할 수 있다.

위의 전처리 과정을 거친 입술(B)와 입술내부(A) 영역에 대한 두 영상을 알맞은 임계값으로 각 영역을 구분한다. 이 후 두 영상을 더함으로서 입술영역만을 검출 할 수 있다.

세 번째 단계, 입술 영역의 픽셀 정보를 이용하여 모양 특징 요소(width, length, width/length, dimension rate, pixel value, length location, width location)를 뽑아낼 수 있다.

3. 음성 오인식률 개선 알고리즘

음성신호는 음성인식기에 의해 분석되어져 DB에 있는 가장 유사한 단어를 얻을 수 있다. 인식된 단어는 아스키 코드를 분석하

여 각 음절의 자음과 모음을 찾아낸다. 추출된 모음과 순서정보는 표1의 알고리즘을 이용하여 비전에서의 모음 데이터와 비교하여 인식된 단어를 거절할지, 채택할지를 판단하게 된다.

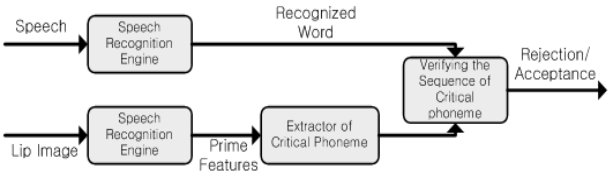

Fig. 3 Flow of an auditory-visual speech recognition to reduce the word error rate

기준데이터를 선정하는 것은 영상에서 추출한 모음 중 가중치가 가장 높은 모음을 찾는 것으로부터 시작된다. 만약 가중치가 같은 모음이 발견 된다면 영상데이터 상에서 앞에 존재하는 모음을 기준데이터로 지정한다. 음성데이터 중에서 선정된 기준데이터와 같은 모음을 찾고, 이 모음의 앞뒤 모음을 확인한다. Fig 4는 음성데이터에서 같은 모음의 존재를 고려한 Tree구조를 나타내고 있다.

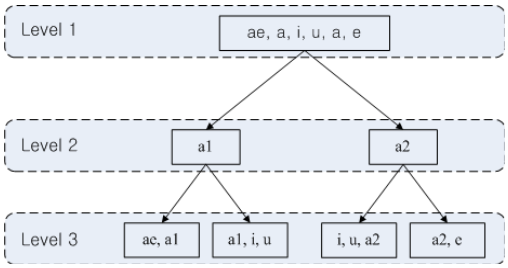

Fig. 4 Tree of the vowel analysis

Tree 구조를 보면 [a]가 2개 존재하기 때문에 Level 2와 같이 [a1], [a2]로 나눠지며, Level 3과 같이 앞뒤에 있는 모음들로 구분 되어진다. 또한 a1뒤와 a2앞에 있는 모음은 중복되기 때문에, 중복되어진 모음에 관해서는 1번만 계산한다. 오차율을 구하기 위하여 Level 3을 이용하며, 표1은 오인식률을 개선하기 위한 알고리즘이다.

Table.1 Algorithm for recognition

```
Begin ()
 STEP1
 if( V1=A(모음데이터)에서 가장 높은 weight를 갖는 모음 추출)
   if(A에 weight가 같은 vowel이 있다면)
      음성 모음순서에서 앞에 있는 것 사용
 else  go to step 2
 end
 STEP2
    if(B에 V1이 중복 되어 있다면 )
       Tree 구조와 같이 표본 추출
  go to step 3
 end
 STEP3
 A와 B의 일치율 계산
 if( 총점이 threshold 보다 클때)  통과
 else  탈락
 end
End (Recognition)
```

[ae]의 맞을 확률이 ε이라 정의하고, 정의된 확률에 따라 추출한 모음이 [ae]라면 1을 곱하고 틀리면 (1-ε)/ε 로 계산한다. 이 후 [a1]의 weight인 α를 곱하여 일치율을 구한다. 여기에서 [a1]과 [ae]가 정확하다면 점수가 1이 되어야 하지만 α를 곱해주는 이유는 [a1]과 [ae]의 가중치 이외에도 모음 순서를 고려하여 계산하였기 때문이다.

각 TREE에서 계산된 값을 더한 후 총 개수(n)로 나눔으로서 일치율을 계산할 수 있다. 나온 일치율은 임계치와 비교하여 클 때는 로봇이 실행할 수 있게 통과되며 낮을 때는 실패가 된다.

4. 실험 결과

음성인식을 위한 음성단어인식기는 상용제품인 Voiceware를 사용하였다. 음성정보만을 이용한 음성인식의 경우 인식대상의 단어가 짧을수록 오인식률이 높다. 따라서 오인식률 개선을 위한 인식대상 단어는 짧은 음절로 선정하였고, 비전데이터에서 추출한 모음데이터에 Weight를 적용하기 위하여 사전에 모음에 대한 반복테스트를 하였다. 표2는 주요 모음에 대한 입술형태와 제시한 알고리즘을 적용하기 위한 특징을 보여준다.

Table.2 Lip shape and the prime features of each critical vowel

	Lip Shape	Width/ Height	Area/ Image Area	Width/ Image width
A		2.51	16.286	67.340
Ae		5.025	10.513	71.276
E		4.957	12.696	48.936
Eo		3.452	9.590	60.638
Eu		6.241	7.306	72.340
I		5.833	9.739	78.723
O		4.431	4.497	48.936
U		4.563	2.604	35.106

영상기반 단어인식 시스템이 음성기반 단어인식기에서 오인식된 단어를 거절함으로서 오인식률이 개선됨을 확인할 수 있었다.

5. 결론

본 논문은 음성기반 단어인식기의 오인식률을 개선하기 위한 방법으로 입술의 움직임에서 각 모음에 대한 주요 특징을 이용하여 모음의 순서를 추출하였다.

각 모음의 추출을 할 때, 단일 프레임 기반으로 모음을 추출하였다. 입술 모양을 이용하여 모음을 보다 신뢰성 있게 추출하기 위해 추후 연구과제로 음성신호를 이용하여 모음을 추출할 수 있는 연속 프레임에 대한 분석이 필요하다.

후기

본 연구는 산업기술연구회 협동연구사업의 연구비 지원으로 수행되었습니다.

참고문헌

1. Woohyun Ko, Sanghoon Ji and Seokwon Lee, "Auditory-Visual Speech Recognition for U-Intelligent Educational Robots", ICMT 2009, October 2009
2. Jong-Seok Lee and Cheol Hoon Park, "Constructing a Noise-Robust Speech Recognition System using Acoustic and Visual Information", Journal of Control, Automation, and Systems Engineering Vol. 13, No. 8, August 2007 pp. 719-725
3. Seong-Young Ko, Jonathan Kim, Woo-Jung Lee and Dong-Soo Kwon, "Surgery Task Model for Intelligent Interaction between Surgeon and Laparoscopic Assistant Robot", International Journal of ARM, Vol. 8, No.1, March 2007 p 38-p46
4. R.H. Taylor, D. Stoianovici, "Medical Robotics in Computer-Integrated Surgery," IEEE Transactions on Robotics and Automation, Vol. 19, No. 5,