2002년도 대한전자공학회/대한전기학회 시스템 및 제어분야 합동 추계학술대회 논문집(2002.11.30)

자율형 이동로봇을 위한 전방위 화자 추종 시스템

이 창 훈, 김 용 호
배재대학교 전자공학과

Speaker Tracking System for Autonomous Mobile Robot

Chang-Hoon Lee, Yong Hoh Kim
Dept. of Electronic Eng., Paichai Univ.
E-mail : naviro@mail.pcu.ac.kr

Abstract

This paper describes a omni-directionally speaker tracking system for mobile robot interface in real environment. Its purpose is to detect a robust 360-degree sound source and to recognize voice command at a long distance(60-300cm). We consider spatial features, the relation of position and interaural time differences, and realize speaker tracking system using fuzzy inference process based on inference rules generated by its spatial features.

I. 서론

전자-컴퓨터 및 기계 기술의 발전으로 로봇은 산업적인 용도뿐만 아니라 가정용, 서비스용, 애완용, 교육용으로까지 그 사용 범위가 넓어져가고 있다. 이러한 사용범위의 확대는 로봇의 활동공간이 특정한 영역에서 점차 인간의 생활공간으로의 확대됨을 의미하고, 이와 같은 인간과 로봇의 활동공간의 공유로 인하여 인간과 접하는 시간이 증대되어 로봇에 있어 대화 기능이 절실히 요구되고 있으며, 이러한 로봇의 대화 기능에서 음성인식이 중요한 역할을 하고 있다. 하지만, 현재 마이크에서 30-60cm 정도의 떨어진 거리에서 음성을 인식하는 것이 일반적이며, 그 이상의 거리에서는 인식률이 급격히 떨어져 실제 이동로봇에 적용하기에는 힘든 상황이다. 그리고 인간은 대화할 때 서로 마주보며 대화하는 것이 일반적이며, 그렇지 않는 경우에는 인간 간의 대화에서도 위화감을 주기 마련이다. 이러한 문제에 대해 하나의 대안으로 화자의 방향을 검지하여 추종하는 것이 있다. 화자의 방향을 검지하여 추종함으로써 로봇과의 대화에 있어 위화감을 줄일 수 있으며, 또한 지향성을 증대시켜 잡음에 대해 상대적으로 신호를 강조할 수 있어 음성인식률을 증대시킬 수 있다.

그래서, 본 논문에서는 대화형 이동로봇에 있어 화자의 방향을 검지하여 추종하는 시스템에 관하여 논한다. 특히, 이동로봇의 특성상 화자와의 거리가 반경 3m이내에서 전방뿐만 아니라 후방, 즉, 임의의 방위에 대하여 화자의 방향을 판별해야하며, 음성인식이 가능해야한다는 제약 조건에서 3개의 마이크로 음성의 입력받아 신호처리, 지연시간차와 음원방향과의 관계에 퍼지이론을 적용하여 화자를 추종하는 시스템을 구현한다.

II. 화자의 방향검지

2.1 시스템 구성

반경 15cm의 정삼각형으로 배열된 3개의 마이크(왼쪽: l, 오른쪽: r, 전방중앙: c)로부터 각각 입력된 신호

를 이용하여 상관관계를 계산하고 마이크 위치의 특성으로부터 얻어진 정성적인 관계를 퍼지추론하여 이를 통하여 음원의 방향을 검출하여 화자를 추종하고, 목적음의 강조와 잡음 제거를 통하여 얻은 결과를 음성인식부로 출력하는 시스템을 구성한다. 시스템의 전체 구성도는 그림 1과 같다.

원거리에서의 음도 받아들일 수 있도록 하기위해 마이크로부터 입력된 신호는 비선형 압축과정을 거쳐 22kHz의 샘플링 주파수로 A/D변환기를 통하여 디지털 신호로 변환된다. 디지털로 변환된 신호는 다시 선형으로 복원되어 방향 검출을 위한 디지털 음원으로 사용된다.

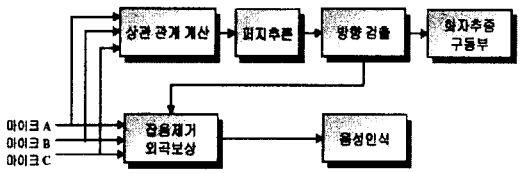


그림 1 전체 시스템 구성도

2.2 방향 검지 알고리즘

방향 판별을 위하여 일반적으로 (i)두 귀사이의 시간 차이(Interaural Time Differences: ITD), (ii)두 귀사이의 음압 차이(Inteaural Intensity Differences: IID), (iii)헤드전달함수(Head-related Transfer Functions: HTF)를 사용한다. 이들 중 두 귀사이의 음압 또는 강도의 차이를 실제로 마이크를 통해 이용할 때는 여러 증폭기의 증폭도를 일정하게 유지하기가 어려우며, 시간에 따른 증폭도 변화 등 유효한 오차 범위를 유지하기 위해서는 여러 가지 어려움이 따른다. 그리고, 두 귀사이의 시간 차이를 이용할 경우 고주파에서는 방향 판별이 어려우나 1.5kHz 근방이나 이하에서는 효율적이다. 본 연구에서는 샘플링 주파수를 높이면 판별의 정확도를 높일 수 있고, 음원이 음성이므로 비교적 간편한 두 귀사이의 시간 차이에 근거한 방향 판별 방법을 고려한다.

두개의 마이크를 이용할 경우 측면 방향에 음원이 있을 경우 판별 오차가 커지며, 전방과 후방을 구별하기가 어렵다. 이러한 문제를 고려하여 모든 방위에 대하여 상당히 균일한 오차를 가지며 전방위를 판별할 수 있게 하기 위하여 본 연구에서는 정삼각형 형태로 3개의 마이크를 배치하여 사용한다.

먼저, 이해를 돕기 위하여 두개의 마이크의 경우에 대하여 살펴보도록 하자.

두 마이크 사이의 거리를 l, 두 마이크를 이은 선분과 수직인 선분과 과 음원과의 각도차를 θ라 하면, 음원에서 두 마이크에 도달하는 거리의 차이 d와의 관계식은 아래와 같다.

$$\theta \approx sin^{-1}\left(\frac{d}{l}\right) \tag{1}$$

$$d = \frac{v}{fs}k \tag{2}$$

여기서, v는 음파의 속도, fs는 샘플링 주파수, k는 위상차이다. 그리고 두 마이크로 입력된 신호를 각각 $s_x(n)$, $s_y(n)$라 하면 두 신호 사이의 상호-상관관계 $R_{xy}(k)$는 다음과 같다. 이 값 중에서 최대가 되는 k가 지연된 샘플의 차이 값이 된다.

$$R_{xy}(k) = \frac{\sum_n \{s_x(n-k)s_y(n)\}}{\sqrt{\sum_n s_x(n-k)^2}\sqrt{\sum_n s_y(n)^2}} \tag{3}$$

이제 3개의 마이크를 원점을 중심으로 정삼각형으로 배치하여, 각 마이크를 통해 입력된 디지털 신호를 $s_l(n)$, $s_r(n)$, $s_c(n)$라 하고, 이들 중 각 쌍의 상관관계, $R_{rl}(k)$, $R_{rc}(k)$, $R_{cl}(k)$라 하자. 이때의 값은 -1에서 1의 값을 가진다.

그러면, 위치와 시간 지연과의 관계를 살펴보자. 반경 60cm-300cm와 각도 1-360도에 해당하는 위치와 왼쪽과 오른쪽 마이크 사이의 지연 시간과의 관계를 계산하여 정규화한 것을 그림 2와 3에 나타낸다. 그림 2와 3을 보면 지연 시간은 거의 각도에 의존함을 알 수 있다. 또한 마이크가 x축과 평행하게 배치되어 있으므로 그림에서 볼 수 있듯이 수평 방향으로 조밀하게 되어있어 분해능이 떨어짐을 알 수 있다.

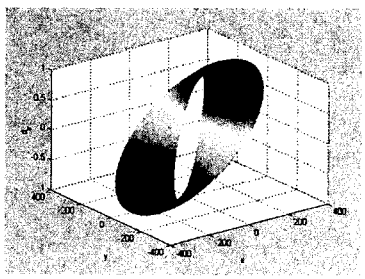


그림 2 위치와 지연 시간과의 관계, $d_{lr}(x, y)$

이에 대해 정삼각형으로 배치함으로써 조밀하게 된 각도 영역에 대하여 그림 4와 같이 또 다른 마이크 쌍에 의해 보완되어지게 된다. 이것이 전방위에 대해 분

해능을 일정하게 유지할 수 있는 이유이다. 그리고 마이크 사이의 간격 정보를 이용하여 최대 지연 시간을 계산하여 이를 이용함으로써 반사음이나 일부 예외 상황에 대한 영향을 배재시키는 효과를 얻을 수 있다.

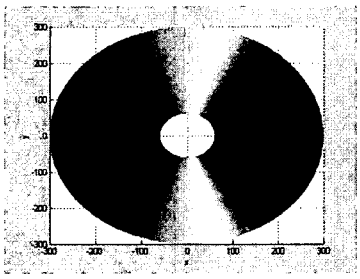


그림 3 위치와 지연 시간과의 관계 (마이크 l과 r)

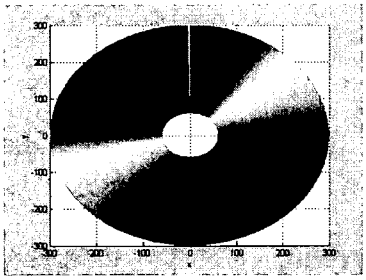


그림 4 위치와 지연 시간과의 관계, $d_{rc}(x,y)$ (마이크 r과 c)

그 다음은 어느 각도에서 화자가 말을 했는지의 여부를 알아내기 위하여, 앞에서 얻은 두개의 마이크 쌍에 대한 위치와 지연 시간과의 관계로부터 각도를 추론할 수 있다.

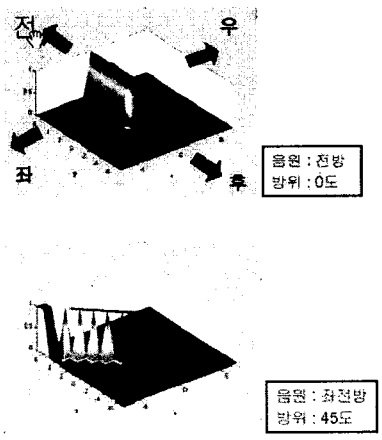
표 1 방위 계산을 위한 퍼지추론 규칙

	Mic l-r	Mic r-c	Mic c-l
30	NM	PB	NM
60	NB	PB	ZR
90	NB	PM	PM
120	NB	ZR	PB
150	NM	NM	PB
180	ZR	NB	PB
210	PM	NB	PM
240	PB	NB	ZR
270	PB	NM	NM
300	PB	ZR	NB
330	PM	PM	NB
360	ZR	PB	NB

예를 들어, 음원이 180도일 때를 생각해 보자. 마이크 l과 r에 대해서는 둘을 잇는 선분과 수직이므로 지연이 없어 0이 됨을 알 수 있다. 그리고 마이크 r과 c에 대해서는 음으로 큰 값을 갖게 된다. 또한 마이크 c과 l에 대해서는 양으로 큰 값을 가짐을 알 수 있다. 반대로 입력 신호로부터 상호-상관관계를 계산하여 지연 시간이 위와 같은 경우에도 서로의 관계를 종합하여 추론하면 방위 값을 얻을 수 있음을 알 수 있다. 이와 같은 관계로부터 30도 간격으로 작성한 추론규칙은 표 1과 같다. 자세한 지연 시간과 각도와의 관계는 이 표의 규칙을 보면 이해할 수 있다.

III. 실험 및 결과 고찰

본 연구에서는 두 귀 사이의 시간차를 근거로 반경 15cm로 3개의 -38dB 감도의 무지향성 콘덴서 3개의 마이크로폰을 정삼각형으로 구성하여, 입력 신호들 간의 지연 시간을 계산하여 각도와 지연 시간과의 관계로부터 얻어진 표 1의 퍼지규칙으로부터 퍼지추론을 하여 전방위에 대해 균일한 분해능으로 화자의 방위를 얻도록 구성하였다. 구성은 그림 1과 같으나, 마이크와 신호처리부와의 사이에는 60-300cm 거리에서도 화자의 방향을 검지하고 음성인식부의 입력으로 사용하여 인식 가능하도록 하기 위하여 신호압축 등의 전처리 회로가 구성되어 있고, A/D변환기로는 National Instrument사의 12비트, 100kHz의 카드형을 이용하였다. 실험에서는 각 채널에 대해 22kHz로 샘플링하였고, 입력신호 범위는 ±2.5V이었다. 음파의 속도는 20도에서 343.4m/s로 하였으며, 이때 최대 지연 샘플은 17이었다. 그리고 퍼지추론 입력으로는 3쌍의 마이크에 대한 상호-상관관계 계수값이 최대가 되는 위상차(지연 샘플수)의 상위 3값을 이용하였다.



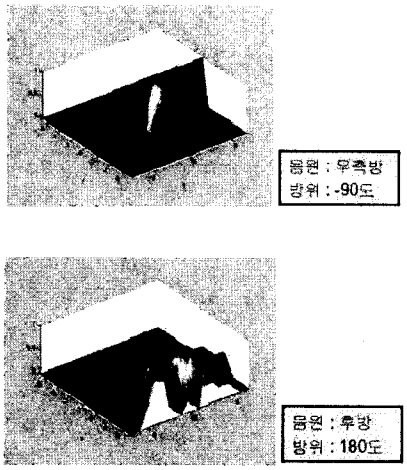


그림 5 0, 45, -90, 180에서의 음성에 대한 추론결과

실험한 결과를 그림 5에 나타낸다. 잡음이 없는 이상적인 경우에는 화자의 방향에 따라 매끄러운 그래프를 그리지만, 잡음이 있는 실제 상황에서는 매끄러운 결과가 나오지 않는다. 그래서 방향 검지에 앞서 일부에 음성 판별을 위한 문턱값을 정하는 알고리즘의 일부를 사용하였다. 잡음의 정확한 레벨은 측정하지 못하였으나, 팬의 모터 잡음과 인간이 충분히 들을 수 있을 정도의 FM방송이 있는 환경에서 실험하였다. 이때의 방위 검지의 오차 범위는 ±5도 이내이었다. 또한 온도가 다소 차이가 나는 환경에서 실험한 결과도 양호하게 나타났다. 이처럼 온도 변화가 있는 환경에서도 오차가 커지지 않는 이유는 정성적인 규칙에 바탕으로 퍼지추론을 하기 때문인 것으로 판단된다.

방위 검지한 후 결과를 음성인식부의 입력으로 사용하여 인식한 결과 60-300cm 거리에서 인식률이 85% 정도이었다. 이것은 근거리와 원거리에 관계없이 인식한 결과로 양호한 것으로 보여진다. 인식률이 다소 떨어진 이유로는 압축된 신호를 완전히 선형으로 복원하지 않아 음성에 왜곡이 생긴 것이 하나의 이유이고, 또 하나는 음성 인식 프로그램 자체의 인식률이 환경에 민감한 것이 이유로 여겨진다.

이번 실험에서는 적용하지 않았지만, 검지된 방향으로 정면을 향하도록 구동부를 작동시켜, 이와 함께 검지된 방위 정보를 바탕으로 화자의 위치로부터 각 마이크에 도달하는 음의 지연시간을 계산할 수 있기 때문에 이를 이용하여 각 신호를 동기화시킬 수 있고, 음원 이외의 신호는 위상이 달라지므로 동기화시킨 신호의 합을 통하여 음원에 대하여 상대적으로 잡음을 줄일 수 있다. 이를 적용하면 높은 감도의 위치에 맞춤으로써 지향성 마이크 역할을 하게 되어 신호 대 잡음비를 높임으로써 음성 인식률을 다소 향상시킬 수 있을 것으로 여겨진다.

Ⅳ. 결론

본 연구에서는 자율이동로봇과 인간과의 인터페이스에 있어 음성을 이용하여 화자의 방향을 검지하여 로봇이 화자를 추종함으로써 위화감을 줄이고, 3m 이내에서 명령한 것을 로봇이 음성인식을 하여 명령을 수행하게 하기위한 목적을 신호처리와 화자의 위치와 신호들의 상관관계를 정성적으로 표현하여 퍼지추론을 통하여 실현하였다.

인간과 같이 두 귀에 해당하는 2개의 마이크를 이용하여 방향 검지를 하는 경우에는 본문에서 언급한 바와 같이 마이크를 잇는 선분에 가까운 곳에서의 음원을 검지할 경우에는 오차가 커질 수밖에 없으며, 전방위에 대한 검지도 어렵다. 그렇지 않는 경우에는 상당히 복잡한 알고리즘으로 구현하여야 하는 부담이 생긴다. 이에 반해 3개의 마이크를 사용하는 경우에는 비교적 간단한 방법으로 전방위에 대해 분해능을 상당히 균일하게 유지할 수 있으며 전방위에 대해 화자의 방향을 검지할 수 있음을 알 수 있다.

대화형 로봇의 가정에서의 실용화를 위해서는 청소기와 같은 강한 유색잡음이 있는 경우에 음성 구분과 TV소리, 다화자의 경우에 화자 구분 등의 기술이 요구된다. 그래서 앞으로 유용한 신호처리 기법의 도입과 화상과의 융합을 고려하고 있으며, 이러한 문제에 대한 알고리즘 개발에 관한 연구를 진행할 것이다.

참고문헌

[1] Jens Blauer, Spatial Hearing : The Psychophysics of Human Sound Localization. MIT Press, 1996.

[2] C. Schauer, H.-M. Gross, "Model and Application of a Binaural 360° Sound Localization System," IEEE Conf. pp.1132-1137, 2001.

[3] N. Roman, D. Wang, G. J. Brown, "Speech segregation based on sound localization," IEEE Conf., pp.2861-2866, 2001.

[4] K. Nakadai, K. Hidai, H. G. Okuno, H. Kitano, "Real-Time Active Human Tracking by Hierarchical Integration of Audition and Vision," Proc. of IEEE-RAS Humanoid 2001, pp.91-98, Nov. 2001.