멀티모달러티를 이용한 실시간 음원추적 시스템 구현

박 정 옥, 나 승 유, 김 진 영
전남대학교 전자정보통신공학과, 전남대 지역협력연구센터

The Implementation of Real-Time Speaker Localization Using Multi-Modality

Jeong ok Park, Seung You Na, Jin Young Kim
Dept. of Electronics Engineering, Chonnam National University, RRC HECS

Abstract - This paper presents an implementation of real-time speaker localization using audio-visual information. Four channels of microphone signals are processed to detect vertical as well as horizontal speaker positions. At first short-time average magnitude difference function(AMDF) signals are used to determine whether the microphone signals are human voices or not. And then the orientation and distance information of the sound sources can be obtained through interaural time difference and interaual level differences. Finally visual information by a camera helps get finer tuning of the speaker orientation. Experimental results of the real-time localization system show that the performance improves to 99.6% compared to the rate of 88.8% when only the audio information is used.

Key Words :음상정위, 시청각 정보

Ⅰ. 서론

인간은 두 귀를 가지고 음이 발생한 방향(상/하, 좌/우)과 거리를 알 수 있다. 동일하게 시스템으로 구현할 경우 인간의 두 귀를 대신한 두 개의 마이크를 사용하면 상/하, 혹은 좌/우 추적 중 하나만을 사용할 수 밖에 없다. 따라서 인간의 귀와 유사해지려면 최소 3개 이상의 마이크를 사용하여야만 3차원적으로 음원을 추적할 수 있다. 그리고 입력된 신호를 가지고 음성의 여부를 판단하기 위한 수단으로 피치를 구한 후 일정 영역에 들어오는지를 확인할 수 있다. 그 다음 마이크에 입력된 음파로 음원의 발생 방향과 거리를 알아낼 수 있다. 주된 방법은 두 마이크 각각에 도달하는 음으로 시간차(Interaural Time Difference: ITD)와 레벨차(Interaural Level Difference : ILD)를 구해 방향과 거리를 알아낸다[1].

얼굴 검출 기술은 시점 변화, 자세 및 조명 변화 등에 존재하는 다양성을 수용하는 알고리즘을 사용하며, 이는 시스템의 처리 속도에 큰 영향을 미친다. 따라서 검출 알고리즘의 처리 속도를 향상시키는 것은 실질적인 응용 시스템을 개발할 때 고려해야 할 중요한 문제이다[2].

본 논문에서는 기 구현된 시스템[3]에 부가하여 상/하 2채널, 피치를 검출 음성의 유무를 판별하여 음성일 경우 음상정위를 판별하도록 구현하였다. 그리고 추적 성능 향상을 위하여 카메라를 통해 입력된 영상 정보를 바탕으로 화자의 위치를 정확히 추정할 수 있도록 모터를 제어 보정하였으며, 이를 실시간으로 구현하였다.

Ⅱ. 음원 정위 시스템

본 논문에서는 4채널(상/하, 좌/우) 마이크로 입력되는 신호의 음성 여부를 판단하여 음원 추적을 구현하였으며, 음원 추적의 보정을 위하여 카메라로 입력되는 영상 신호를 이용 얼굴 영역을 검출하여 보다 정확한 음원 정위를 추정하였다.

1. 음성의 주기성을 이용한 음성여부 결정

본 논문에서 구현된 피치검출 프로그램은 음성 데이터가 들어오면 각 프레임 별로 에너지를 추출하는데 그 에너지가 최대가 되는 프레임을 찾아 그 정보를 가지고 short-time average magnitude difference function(AMDF) 신호를 이용하여 아래 그림 1과 같이 null points(AMDF신호에서 꼭지점들)를 찾아서 범위 결정을 하였으며, 다음 식 1은 $AMDF(\gamma_n(k))$의 정의 식이다.

$$\gamma_n(k)=\sum_{m=-\infty}^{\infty}|x(n+m)w_1(m)-x(n+m-k)w_2(m-k)| \quad (1)$$

여기서 $x(n)$은 입력신호이고 $w_1(m)$, $w_2(m)$은 윈도우들이고 k는 delay이다. 그리고 음성여부를 판단한 피치의 범위(70Hz~350Hz)는 학습데이터의 실험으로서 결정되었으며 실험에서는 음성인 데이터만을 가지고 음원 정위를 구하였다

2. 음원 정위에 사용된 알고리즘

1절에서 제시한 음성의 유무를 판별한 다음 음성일 경우 음원이 발생한 지점을 추정하기 위해서 음원을 특정 각도에 정위시켜야 한다. 이 과정을 수행하기 위한 가장 중요한

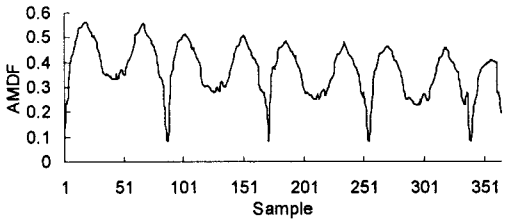


그림 1. AMDF 처리 결과의 예시

저자 소개
* 朴 正 玉 : 全南大學 電子情報通信工學科 碩士課程
** 羅 承 裕 : 全南大學 電子情報通信工學科 敎授・工博
***金 秦 永 : 全南大學 電子情報通信工學科 敎授・工博

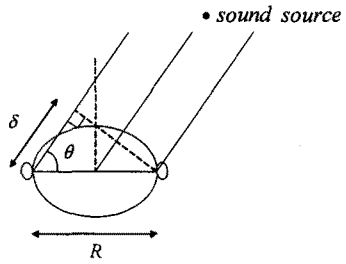


그림 2. 음원 정위의 원리

단서는 수평면에 위치한 두 귀에서 파면의 상대적인 차이, 즉 두 귀에 입사하는 신호의 차이를 이용한다. 아래의 그림 2는 음원 발생한 각을 추정하기 위한 것이다. 여기서 입사음은 평면파라고 가정을 한다. 하지만 음성은 구면파이기 때문에 오차가 발생하게 된다. 또한 마이크로 입력받은 데이터는 마이크의 전기적인 특성으로 인한 잡음과 외부의 노이즈도 함께 들어와 각을 추정하기에 상당한 애로사항이 발생한다.이 점을 보완하기 위하여 2.2KHz 저역 통과 필터링을 하였다.

그림에서 일정한 간격(20cm)으로 놓여진 마이크를 통하여 입력된 4채널(상/하, 좌/우)의 음성신호($x(t)$, $r(t)$)를 일반화된 상호상관(cross-correlation)값으로 계산하여 최대 상관도를 보이는 지점을 검출한다. 식(2)는 상호상관 값을 구하기 위한 식이다

$$R_{xr}=E_t[x(t)r(t-\tau)] \quad (2)$$

입력신호의 총 데이터 사이즈와 시간을 이용하면, 거리(δ)를 음파가 전달되는데 걸리는 시간을 구할 수 있다. 음속(C)은 340m/s 이므로 식(3)에 대입하면 음원의 발생 방향(θ)을 알아낼 수 있다[1].

$$\cos\theta\approx\frac{\delta}{R} \quad (3)$$

3. 음원정위 보정위한 얼굴 영역 검출 알고리즘

음원 추적 과정에서 마이크에 입력된 음의 굴절, 반사, 회절로 인한 계산 오차를 보완하기 위하여 음원 추적 후 얼굴 영역 검출 알고리즘을 사용하여 추적 성능을 높일 수 있다. 얼굴 영역 검출 알고리즘은 학습된 계층적 분류기를 통해, 빠르고 효율적으로 얼굴 영역을 찾아낸다. 검출 알고리즘은 전처리, 실시간 얼굴 영역 검출의 두 단계로 구성된다[4].

실시간 얼굴 영역 검출은 입력된 영상 정보를 사용, summed-area table(SAT)를 생성하고 사각형 마스크로 구성된 계층적 분류기를 이용하여, 전처리 과정에서 생성된 입력 패턴을 분류한다. 분류기는 AdaBoost 알고리즘을 적용하였으며, 얼굴 패턴의 특징을 추출하는데 결정적인 역할을 한다[2].

추적 알고리즘은 서보모터를 통해 동적으로 검출 영역을 확장시키며, 실시간으로 추출된 얼굴 영역의 위치 정보만을 이용함으로써 시스템의 처리 속도를 증가시킨다. 구현된 추적 알고리즘은 분류기 학습, 실시간 얼굴 검출, 얼굴 추적의 세 단계로 구성된다. 여기서 얼굴 검출 단계는 생성된 계층적 분류기를 이용하여, 실시간으로 얼굴 영역을 찾아낸다.

계층적 분류기는 특징들을 이용하여, 입력 패턴에 대한 특징 값을 계산한다. 따라서, 입력 패턴은 계산된 특징 값이 임계값을 만족하면, 얼굴 영역으로 분류된다. 마지막 단계인 얼굴 추적 단계에서는 얼굴 영역의 위치 정보를 이용하여 서보모터를 제어한다.

최종적으로 구현된 음원 정위 시스템의 블록 다이어그램은 그림 3과 같다.

III. 실험 방법

실험은 20대의 남성화자와 여성화자 5명으로 화자의 위치는 10, 45, 90, 135, 170도에 자리하게 하였으며 마이크와 화자의 간격을 1m로 위치하였다. 화자는 일정한 순서에 의해 "무궁화 꽃이 피었습니다."라는 문장을 1회 발음하게 하였다. 사용된 시스템은 4채널 오디오 입/출력 카드와 일반적으로 사용하는 카메라로 구현되었으며, 구현된 시스템의 구성은 아래 그림 4와 같다.

음성입력은 16bit, 44.1KHz로 샘플링되며 입력된 신호의 피치를 구하여 음성 여부를 판단 음성신호만을 사용하였으며 영상입력은 172x144의 이미지를 사용하였다. 실험시 주위환경은 잡음(실험장비의 기계적 노이즈)이 존재하는 사무실 환경에서 실험을 실시하였다.

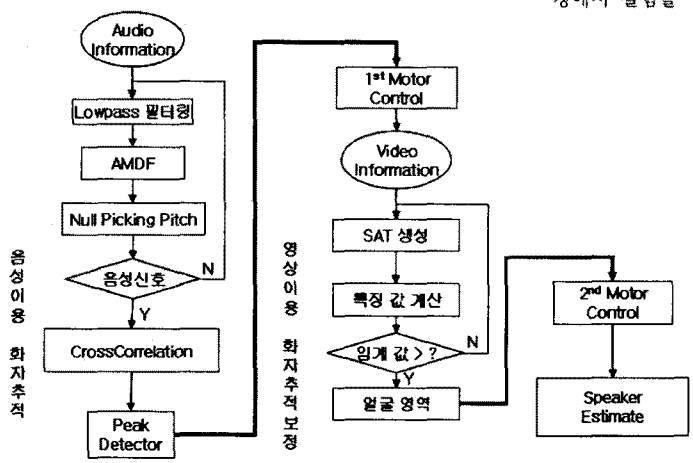


그림 3. 구현 시스템 블록 다이어그램

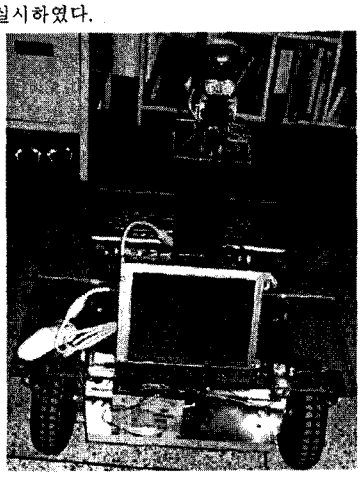
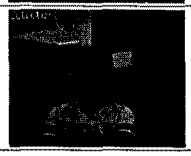


그림 4. 구현 시스템

표 1. 음성/영상신호를 이용 화자 추적 결과

	출력영상
화자 1차 발음 후 시스템이 정확한 추정을 하지 못하여 오차가 발생한 경우	
화자 1차 발음 후 시스템이 정확히 추정 화자가 화면 정중앙에 위치한 경우	
1차 화자 발음 후 각을 추정하여 위치를 찾은 후 영상을 이용 보정한 경우	

실험은 음성신호만을 이용한 상/하, 좌/우 화자 추적과 음성신호와 영상신호를 이용하여 음원정위를 보정한 화자 추적으로 구분하여 실험하였다. 본 실험에서 오차범위는 ±5° 이내로 하였으며 범위를 벗어났으나 화면에 화자의 얼굴이 표시되는 경우는 위치에 따라 ±5의 패널티를 부여하였다.

Ⅳ. 실험 결과

본 실험은 기 구현된 시스템에서 실험한 것에서 새롭게 첨가된 음성의 유무 판단의 정확성, 상/하 마이크 입력을 이용한 각 추정실험을 부가하였다. 마지막으로 동일한 조건에서 사람을 대상으로 위치추정을 함으로써 구현된 시스템과 비교하였다.

1. 음성유무의 판단 실험결과

음성 외에 박수소리와 같은 고주파 신호를 50회 씩 10회 발생시킴으로서 시스템의 동작여부를 관찰하였다. 그 결과 박수소리와 같은 고주파 신호에는 100%반응하지 않음을 확인할 수 있었다.

2. 상/하, 좌/우 음성 신호 추적 실험 결과

표 1의 첫 번째 그림과 같이 음성만을 사용하여 추적할 결과 오차가 발생하는 것을 볼 수 있으며, 두 번째 그림과 같이 정확히 추정되는 것도 확인할 수 있다. 이러한 오차는 그림 5에서와 같이 오차 값이 5도 이상임을 확인할 수 있는데 좌/우 보다 상/하의 오차 값이 큰 이유는 마이크가 놓여 있는 테이블에서 생기는 음의 반사로 발생되어진다고 추정된다.

3. 음성 & 영상 신호 추적 실험 결과

음성만을 이용한 실험에서 발생한 오차를 보정하기 위하여 음성과 영상을 함께 사용하여 화자 추적한 결과를 표1의 세 번째 그림에서 나타내었다. 오차를 보정하기 위해 화자의 얼굴 영역을 검출하여 얼굴 영역이 화면 중앙에 위치할 수 있도록 모터를 움직여 카메라의 중앙에 화자가 정확히 들어오도록 하였다. 그림 5에서 알 수 있듯이 음성신호만을 이용한 결과보다 오차 값이 0에 수렴하여 성능 향상을 볼 수 있다.

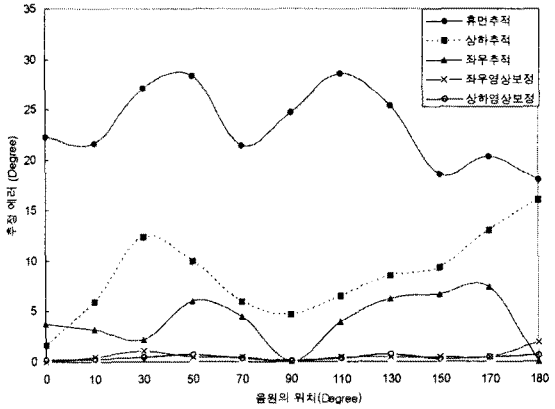


그림 5. 화자 추적 결과

4. 휴먼 추적 실험

구현된 시스템의 성능을 비교하기 위하여 사람을 대상으로 특정 위치에서 녹음된 음성을 들려주어 방향을 판별하도록 하였다. 청각이 예민한 사람의 경우 정확한 각을 가리켰지만 그렇지 못한 사람의 경우는 상당히 틀린 각을 가리켰다. 결과는 사람의 개별 특성으로 인한 편차가 심하여 오차가 30도 이상 발생하였으며 실험결과는 그림 5에서 확인할 수 있다.

Ⅴ. 결론

본 논문에서는 기 구현 시스템[3]에서 새롭게 피치검출 부분을 첨가하여 음성에만 반응하며, 상/하 추적이 가능하도록 변경하였다. 하지만 음원 추적과정에서 발생하는 잡음, 방향차로 인한 추적 오차를 줄이지 못하였으며 이 오차는 영상을 통한 얼굴 추적으로 성능을 향상 시켰다. 향후 음성 신호를 사용한 화자 인증 및 단어 인식 그리고 영상 신호를 이용한 얼굴인식 등의 연구가 필요하다.

감사의 글

본 연구는 한국과학재단 지정 전남대학교 고품질 전기전자부품 및 시스템 연구센터의 연구비 지원에 의해 연구되었음.

참 고 문 헌

[1] C. Schauer, H.-M. Gross, "Model and application of a binaural 360° sound localization system," in Proceedings of the International Joint Conference on Neural Networks, Vol. 2, pp. 1132-1137, 2001.

[2] P. Viola, M. Jones, "Robust real-time face detection," in Proceedings of International Conference on Computer Vision, Vol. 2, pp. 747-747, 2001.

[3] 박정옥, 나승유, 김진영, "휴머노이드 로봇을 위한 시청각 정보 기반 음원 정위 시스템 구현," 제 17회 신호처리합동 학술대회, pp. 84, 2004.

[4] 김수희, 이배호, "계층적 분류기를 이용한 실시간 얼굴 검출 및 추적," 한국정보처리학회 추계학술발표대회, 제10권, 제2호, pp.137-140, 2003.