하모닉 구조를 이용한 두 명의 동시 발화 화자의 위치 추정

김현경 임성길 이현수
경희대학교 컴퓨터공학과

Two Simultaneous Speakers Localization using harmonic structure

Hyunkyung Kim, Sungkil Lim, Hyonsoo Lee
Department of Computer Engineering
KyungHee Univ.

izeit79@empal.com, skan7203@empal.com, leehs@khu.ac.kr

Abstract

In this paper, we propose a sound localization algorithm for two simultaneous speakers. Because speech is wide-band signal, there are many frequency sub-bands in that two speech sounds are mixed. However, in some sub-bands, one speech sound is more dominant than other sounds. In such sub-bands, dominant speech sounds are little interfered by other speech or noise.

In speech sounds, overtones of fundamental frequency have large amplitude, and that are called 'Harmonic structure of speech'. Sub-bands in harmonic structure are more likely dominant. Therefore, the proposed localization algorithm is based on harmonic structure of each speakers. At first, sub-bands that belong to harmonic structure of each speech signal are selected. And then, two speakers are localized using selected sub-bands. The result of simulation shows that localization using selected sub-bands are more efficient and precise than localization methods using all sub-bands.

Ⅰ. 서론

동시 발화 화자의 위치 추정은 다수의 화자가 동시에 발화하는 환경에서 각 화자의 위치를 추정하는 문제로 지능 로봇의 청각시스템, 화상회의 등에 응용할 수 있고 음성분리의 전처리로 사용될 수 있다.[1]

이러한 문제를 해결하기 위하여 Bank는 먼저, 두 개의 마이크로폰으로부터 입력받은 신호를 DFT처리하여 서브밴드로 분할한다. 서브밴드로 나누어진 신호의 오른쪽과 왼쪽의 위상차를 이용하여 거리차이(path difference)를 계산한다. 모든 서브 밴드에서 계산된 거리차이가 동일하지 않기 때문에 가중치 윈도우를 사용하여 두드러진 2개의 거리차이를 선택하고 거리차를 이용하여 방향을 추정한다.[1] Wang은 ERB filter bank를 사용하여 신호를 서브밴드로 분할하고 오른쪽과 왼쪽 각 서브밴드별로 cross correlation을 계산한다. 피크가 존재하는 봉우리를 가우시안 함수를 사용하여 정규화 시키고 모든 서브밴드의 cross correlation 결과를 합한다. 합한 결과에서 가장 큰 피크를 갖는 위치 정보를 이용하여 하나의 방향에서 나타나는 시간지연(ITD)을 계산한다. 첫 번째로 선택된 시간지연과 동일한 시간지연을 갖는 서브밴드들을 제거하고 나머지 서브밴드의 cross correlation의 결과를 합해 가장 큰 피크를 갖는 위치 정보를 이용하여 또 다른 방향에서 나타나는 시간지연을 계산하고 한다. 계산된 시간지연을 방위각으로 맵핑시켜 화자의 위치를 추정한다.[2][3]

기존의 방법들은 신호를 모든 서브밴드를 사용하여 화자의 위치를 추정한다. 서브밴드들 중 한 화자가 두드러진 서브밴드를 사용하면 정확한 위치를 추정할 수 있지만 두 화자가 비슷한 크기로 섞여있는 서브밴드를 사용하여 위치를 추정하면 오차가 발생하기 때문에 정확한 위치를 추정 할 수 없다. 그러므로 기존의 방법처럼 모든 서브밴드를 사용하여 위치를 추정하는 것보다 한 화자의 음성이 두드러진 서브밴드들을 사용하여

위치를 추정하는 것이 정확성을 높이고 계산량도 줄일 수 있다.

하모닉 구조는 기본주파수의 정수배가 되는 주파수를 의미하며 에너지(amplitude)가 크게 나타나다. 에너지가 크기 때문에 다른 화자나 잡음에 의한 영향을 적게 받는다. 일반적으로 하모닉 구조를 갖는 서브밴드에서 한 화자의 음성이 두드러지게 나타난다. 그러므로 본 논문에서는 다른 화자나 잡음에 영향을 적게 받는 하모닉 구조를 이용하여 동시 발화 화자의 위치 추정 방법을 제안한다. 음성을 입력받기 위하여 두 개의 마이크로폰을 사용한다. STFT(Short Time Fourier Transform)변환하고 위치를 추정하기 위한 단서로 시간지연을 이용한다. 시간지연은 채널별 위상차를 이용하여 계산하고 화자의 위치는 방위만 추정한다.

논문의 구성은 2장에서는 제안하는 방법, 3장에서는 실험 및 분석에 대하여 설명하고 마지막으로 4장에는 결론 및 향후연구에 대하여 기술 한다.

Ⅱ. 하모닉 구조를 이용한 두 화자의 위치 추정

본 논문에서는 하모닉 구조를 이용하여 두 화자의 위치 추정하는 방법을 제안한다. 시스템은 크게 주파수 분석, 채널선택, 시간지연 계산, 방위각 판단의 4부분으로 구성된다. 제안하는 시스템 구조는 그림1과 같다.

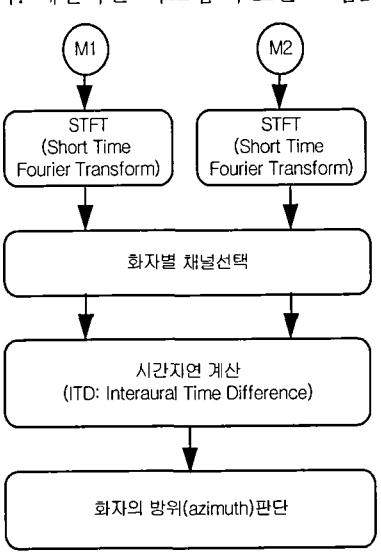

그림 1 제안하는 시스템 구조

두 개의 마이크로폰을 통해 입력받은 오른쪽, 왼쪽 신호는 STFT변환을 한다. 채널 선택부에서는 각 화자별로 오른쪽과 왼쪽의 하모닉 구조를 갖는 채널들을 선택한다. 시간지연 계산부에서는 각 화자별로 선택된 오른쪽과 왼쪽 채널들의 위상 차이를 이용하여 시간지연을 계산한다. 마지막으로, 화자의 방위각 판단 부분에서는 각 화자별로 구한 시간지연과 음파의 속도와의 관계를 이용하여 화자의 방위각을 추정한다.

1. 화자별 채널 선택

혼합된 신호에서, 한 화자의 에너지가 다른 화자의 에너지 보다 매우 큰 채널(이후 도미넌트 채널 이라 함)은 다른 화자나 잡음에 의한 영향을 적게 받는다. 음성의 경우에는 기본 주파수의 정수배가 되는 채널에서 높은 에너지를 갖는 하모닉 구조를 이루고 있는데, 하모닉 구조에 해당하는 채널들은 높은 에너지를 가지고 있기 때문에 도미넌트한 채널이 될 확률이 높다. 따라서 본 논문에서는 하모닉 구조를 이용하여 다른 화자나 잡음에 의한 영향을 적게 받는 채널을 선택하여 방향을 추정한다.

각 화자의 도미넌트한 채널을 선택 하기 위하여 피치 후보 생성, 생성된 후보들의 피치가 될 확률 계산, 하모닉 구조를 갖는 채널을 선택하는 3단계의 처리 과정을 거친다. 피치 후보 생성과 후보의 피치가 될 확률 계산 단계에서는 Kwon, Y.-H 가 제안한 두 화자의 피치 추정 방법[4]의 일부를 제안한 시스템에 맞게 수정하여 사용한다.

먼저, 피치 후보를 생성하기 위하여 frame의 파워스펙트럼을 구한다. 피치가 존재하는 60~280Hz사이에서 피크들을 선택하고 각 피크가 속한 봉우리의 시작점과 끝점(미분하면 이상적으로 0이 되는 위치)을 구한다. 선택된 피크의 양 끝점에 m배를 하고, 그 구간 안에 피크가 존재하면 m번째 하모닉 구조를 이루는 피크로 판단하고 m으로 나누어 피치후보를 생성한다. 양끝 점의 주파수도 선택된 피크의 양끝 점의 주파수를 m으로 나눈 값으로 설정한다. 그림 2는 피크와 양끝 점의 주파수 값을 이용하여 피치후보를 찾는 예를 보여준다.

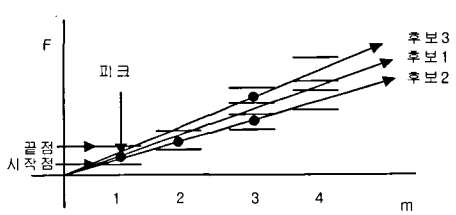

그림 2 피크를 사용하여 피치후보 3개를 생성한 예(F: 주파수, m)정수배, 시작점, 끝점: 피크가 속한 봉우리의 양끝점의 주파수)

각 화자의 하모닉 구조에 해당하는 채널들을 선택하기 위하여 후보 피치들의 배음 채널이 피크가 될 우도(Likely Hood)를 정의한다. 우도를 정의하기 위하여 피크들을 정규화한 값($F(x)$)을 사용하였으며 식2와 같이 정의한다.

$$F(x)=\begin{cases} 1 & \text{if } amp(x)>amp(x-1) \text{ and } amp(x)>amp(x+1) \\ \dfrac{amp(x)}{amp(x+1)} & \text{if } amp(x+1)>amp(x) \text{ and } amp(x+1)>amp(x+2) \\ \dfrac{amp(x)}{amp(x-1)} & \text{if } amp(x-1)>amp(x) \text{ and } amp(x-1)>amp(x-2) \\ 0 & otherwise \end{cases} \quad (식\ 2)$$

식2에서 x는 채널을 나타내며, $amp(x)$는 x채널의 스펙트럼 크기를 의미한다.

피치 후보가 N개 일 때 n번째 피치후보가 피치가 될 우도 $L(n)$은 식3과 같이 정의 한다.

$$L(n)=\frac{1}{J}\sum_{j=1}^{J}F(j\cdot c(n)) \quad where \quad J=\left\lfloor \frac{\theta}{c(n)} \right\rfloor \quad (식\ 3)$$

$c(n)$은 n번째 후보가 속한 채널이고 $j\cdot c(n)$은 $c(n)$의 j번째 하모닉 채널, θ는 피크가 존재하는 범위를 넘지 않는 정수를 의미한다. 본 논문에서 θ는 2kHz를 사용하였다. 두 화자의 피치후보의 우도가 최대가 되는 n_1을 첫 번째 화자의 피치로 결정하고 두 번째 화자의 피치 n_2는 식 4의 조건을 만족하는 피치 후보를 제거하고 나머지 피치 후보들 중 우도가 최대가 되는 것을 선택한다.

$$|c(n_1)-c(n_2)|<R \quad (식\ 4)$$
$$c(n_2)=2(n_1)$$

식 4에서 R은 한 채널이 나타내는 주파수의 범위를 나타낸다.

n_1, n_2가 선택된 두 개의 피치 후보일 때 각 화자의 선택된 채널은 식 5와 같다..

$$Ch(n_1)=\left\{ j\cdot c(n_1) \middle| \begin{array}{l} F(j\cdot c(n_1))\neq 0 \\ \text{and } j\cdot c(n_1)\neq i\cdot c(n_2) \end{array} \right\} \quad (식\ 5)$$
$$\forall\, 1<j<\left\lfloor \frac{\theta}{c(n_2)} \right\rfloor, where\ 1<j<\left\lfloor \frac{\theta}{c(n_1)} \right\rfloor$$
$$Ch(n_2)=\left\{ j\cdot c(n_2) \middle| \begin{array}{l} F(j\cdot c(n_2))\neq 0 \\ \text{and } j\cdot c(n_2)\neq i\cdot c(n_1) \end{array} \right\}$$
$$\forall\, 1<j<\left\lfloor \frac{\theta}{c(n_2)} \right\rfloor, where\ 1<j<\left\lfloor \frac{\theta}{c(n_1)} \right\rfloor$$

식 5에서 첫 번째 조건은 배음채널이 피크채널이거나 이웃채널이 피크 채널이면 n_1의 채널로 선택한다는 것을 의미한다. 두 번째 조건은 화자별로 선택된 n_1, n_2 피치의 동일한 배음채널은 제외시킨다는 것을 의미한다.

2. 시간지연(ITD) 계산

각 화자 별로 채널을 선택한 다음에 선택된 오른쪽과 왼쪽 채널의 위상차를 이용하여 시간지연을 계산한다. 먼저 피치 n_1의 선택된 채널인 $Ch(n_1)$에 속하는 m 번째 채널의 위상차는 식 6과 같이 구한다.

$$\triangle t_m=\frac{\phi_m^l-\phi_m^r}{2\pi f_m} \quad (m\in Ch(n_1)) \quad (식\ 6)$$

식 6에서 $\triangle t_m$은 $Ch(n_1)$에 속하는 m번째 채널의 시간지연, $\phi_m^l-\phi_m^r$은 m번째 채널의 위상차, f_m은 m번째 채널의 주파수를 의미한다.

$Ch(n_1)$의 모든 채널의 위상차가 동일하지 않고 오차를 가지고 있기 때문에 피치 n_1의 시간지연은 채널별로 구한 시간지연의 평균으로 결정한다. 피치 n_2의 시간지연도 n_1의 시간지연을 구하는 방법과 동일하게 구한다.

3. 화자의 방위각 판단

음원의 방위를 판단하기 위한 방법은 음파의 속도와 시간지연을 이용하여 계산하는 방법과 HRTF(Head Related Transfer Function)를 이용한 방법이 있다. HRTF는 고정된 음향 환경에서 관측된 시간지연과 방위각을 사용하기 때문에 화자의 위치를 정확하게 찾을 수 있다는 장점을 갖지만 음향 환경이 변하면 다시 측정해 주어야하는 단점을 가지고 있다. 반면 음파의 속도와 시간지연을 이용하면 HRTF에 비하여 위치 판단의 정확성을 떨어지지만 구현이 쉽고 음향 환경이 변하더라도 사용할 수 있다는 장점을 가지고 있다. 본 논문에서는 음파의 속도와 시간지연을 이용하여 두 화자의 방위를 판단한다. 음파의 속도와 시간지연의 관계는 식 7과 같다.

$$\alpha\approx\cos^{-1}\frac{340\triangle t}{d} \quad (식\ 7)$$

α : 음원의 방위각, d : 마이크로폰 사이의 거리
$\triangle t$: 시간지연, $340m/s$: 음파의 속도

보다 정확한 방향을 추정하기위하여 음원이 빠른 시간 내에 움직이지 않는다는 가정 하에 일정 시간 동안 추정한 방위각의 평균값으로 결정한다.

Ⅲ. 실험 및 분석

제안한 시스템의 동작 검증을 위하여 전체 채널 및 화자별로 선택된 채널의 방위각을 추정하여 비교하는 실험을 하였다. 또한 1초 동안의 데이터를 사용하여 화자의 위치 추정하는 실험을 하였다.

두 화자 중 한 화자는 30°, 다른 화자는 60°에 있고, 마이크로폰 사이의 거리는 27cm, 음성은 1m에서 발생한다고 가정하였다. 실험 데이터는 '아' 발음을 22kHz 샘플링으로 녹음하여 사용하였다.

첫 번째 실험은 2kHz이하의 모든 채널들에서 추정한 방위각과 제안한 방법인 하모닉 구조에 해당하는 채널들에서 추정한 방위각을 비교하였다. 방위각을 10°씩

의 구간으로 나누어 화자의 위치를 추정한다. 그러므로 약 -5° ~+5°의 오차를 가지고 있다. 실험 결과는 그림3과 같다. a)는 2kHz이하의 모든 채널에서 추정한 방위각의 히스토그램이다. 실제 두 화자가 존재하는 30°와 60°이외의 방위각을 나타내는 채널의 수가 많은 것을 볼 수 있다. 음성이 넓은 대역 신호이기 때문에 혼합된 신호에서는 한 화자의 음성이 두드러진 채널수보다 두 음성이 섞여있는 채널수가 많이 존재하기 때문에 발생되는 문제이다. b)는 제안한 방법으로 2kHz 이하의 하모닉 구조에 해당하는 채널만을 사용하여 추정한 방위각의 히스토그램이다. 하모닉 구조를 갖는 채널은 한 화자의 음성이 두드러지게 나타날 확률이 높기 때문에 두 화자가 섞여있는 채널은 선택에서 대부분 제외되었음을 확인 할 수 있다.

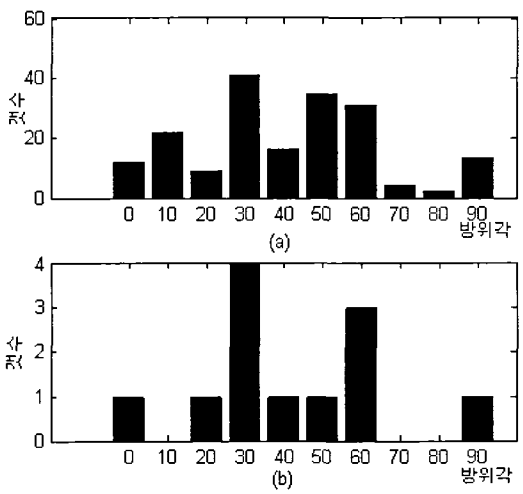

그림 3 전체 채널 및 화자별 선택된 채널에 의한 방위각 계산 비교 (a) 2kHz이하의 모든 채널에서 추정한 방위각의 히스토그램, (b) 제안한 방법으로 2kHz이하의 하모닉 채널에서 추정한 방위각의 히스토그램

두 번째는 약1초 동안의 데이터를 사용하여 화자의 위치 추정하는 실험을 하였다. 결과는 그림 4는 와 같다. a)는 하나의 프레임에서 추정한 두 화자의 방위각을 나타낸다. 두 화자가 위치한 30°와 60°를 기준으로 각도의 오차를 가지고 있다. 이러한 오차를 제거하기 위하여 일정 시간동안 추정된 방위각을 고려하였다. 0.5초 단위로 위치를 추정하였고 결과는 b)와 같다. 위치를 추정한 결과 하나의 프레임에서 나타났던 방위각의 오차를 많이 줄일 수 있었다.

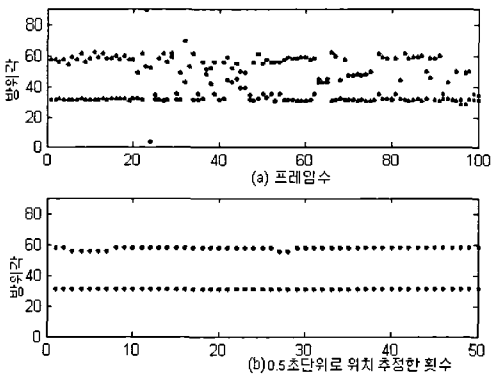

그림 4 하나의 프레임 및 0.5초 단위에서 추정된 방위각 비교 (a) 하나의 프레임에서 추정된 두 화자의 방위각, (b) 0.5초 단위로 위치를 추정한 결과

Ⅳ. 결론 및 향후 연구

본 논문에서는 하모닉 구조를 이용해 두 화자의 위치를 추정하는 방법을 제안하였다. 실험 결과 하모닉 구조에 해당하는 채널은 한 화자의 음성이 도미넌트하게 나타나기 때문에 적은 채널을 사용하더라도 위치를 비교적 정확하게 추정할 수 있다. 하나의 프레임에서 위치를 정확하게 추정하지 못하더라도 일정구간에서 추정된 방위각을 고려하면 해결할 수 있다. 또한, 하모닉 구조를 갖는 적은 채널만을 선택했기 때문에 계산량이 적어 실시간 처리를 하는 응용에서도 사용이 가능하다.

향후에는 제안한 방법의 성능 평가를 위하여 기존의 방법과의 비교가 필요하다.

참고문헌

[1] T Nakatani, M Goto, H.G Okuno, "Localization by harmonic structure and its application to harmonic sound stream segregation", Acoustics, Speech, and Signal Processing, pp.653 - 656, 1996

[2] Banks. D, "Localization and separation of simultaneous voices with two microphones", Communications, Speech and Vision, IEE Proceedings, pp.229-234, 1993

[3] Brian C.J. Moore, "Hearing: Handbook of Perception and Cognition Second Edition" Academic press, pp.347-386, 1995

[4] N. Roman, Wang. DeLiang, G.J. Brown, "Speech segregation based on sound localization", Neural Networks,2001.Proceedings.IJCNN'01. International Joint Conference on, pp.2861-2866, 2001

[5] Y.-H Kwon, D.-J Park, B.-C Ihm, "Simplified pitch detection algorithm of mixed speech signals", Circuits and Systems, pp.722-725, 2000