빔포밍을 이용한 다방향 동시 음성인식

김혜진, 윤호섭, 이재연, 윤영우, 김재홍
한국전자통신연구원 인지기술연구팀
marisan@etri.re.kr, yoonhs@etri.re.kr

Multi-directional Simultaneous Speech Recognition based on Beamforming

Hye-Jin Kim, Ho-sub Yoon
Electronics and Telecommunications Research Institute

요 약

여러 사람이 서로 다른 단어를 발성할 때, 기존의 방법으로는 이 단어들에 대한 음성인식이 어려웠다. 그러나, 화상회의 혹은 로봇 등 마이크로폰을 여러 명이 공통으로 사용하는 환경이 많아지면서 다방향 동시 음성 인식에 대한 요구가 늘어났다. 본 논문은 이를 위해 다방향으로 빔포밍을 적용시켜 동시에 음성인식을 수행하였고, 실시간으로 이 과정들이 수행하도록 하였다. 이 방법은 빔포밍 된 방향에서 올바른 음성인식 단어가 들어오지 않을 때는 이를 잡음으로 간주하도록 한다.

1. 서 론

지능형 서비스를 위한 음성인식에 대한 요구는 날로 증가하고 있다. 또한, 화상회의나 지능형 서비스 로봇과 같이 여러 사람이 모바일 플랫폼을 서로 공유하게 된다. 이러한 상황에서 서로 다른 서비스를 동시에 요청할 수 있도록 하는 기술의 요구 또한 날로 증가하고 있다.

또한 음성 인터페이스 활용의 주제인 잡음 제거도 주된 관심사가 되어왔다. 음원 분리나, Wiener filter와 같은 방법들의 접근들이 많이 있어왔다. 그러나 음원분리는 음원들이 서로 statistically independent하고 음원은 단일음원으로 고유의 특성이 있어야 하며 TV 소리와 같이 수시로 바뀔 수 있는 음원을 다른 소리와 분리하기는 어렵다는 단점이 있다. Wiener filter는 널리 사용되는 잡음제거를 위한 filter로서 stationary noise, 특히 white noise에 강인하나 다양한 특성을 갖는 잡음을 제거하기에는 어렵다. 이를 극복하기 위해 최근에 각광받고 있는 방법으로는 beamforming이 있다. Beamforming은 target signal이 들어오는 방향은 target signal이 강하고 그 외의 잡음은 다른 방향에서 들어온다는 가정을 하게 된다. 여러 개의 microphone을 사용하여 target 방향으로 signal을 steering한다. 이 방법을 통해 target signal은 증폭되고 그 외에 잡음신호들은 간섭이 되어 SNR비를 증대시킬 수 있는 방법으로 사용된다.

2. 관련 연구

2.1 빔포밍

빔포밍은 일반적으로 Generalized Sidelobe Canceller (GSC) 방법과 LCMV[1]방법이 일반적으로 널리 쓰이고 있다. 두 방법 모두 frequency domain과 time domain에서 활용 가능한데, GSC의 큰 장점은 Fixed beamformer (FBF)와 Blocking Matrix(BM), 그리고 Multiple Input Canceller (MIC) 로 구성되어 있다. 특히, BM단과 MIC단은 Adaptation Mode Control (AMC)로 분류되어 FB 단을 통해 가상의 target signal을 얻고 AMC 단계에서 target signal 대비 잡음을 제거하여 빔포밍의 성능을 늘릴 수 있다는 점이다.

빔포밍에서 중요한 것은Fixed beamformer 단계에서 steering이 잘 될 수 있도록 하기 위해 microphone 간의 정확한 time delay와 gain, 그리고 마이크로폰 배열에 따른 signal의 transfer function을 정확하게 구하는 것이다. 이를 수식으로 표현하면 다음과 같다.

$$J = E\left[\left|y_{GSC}(n)\right|^2\right] \tag{1}$$

$$y_{GSC}(n) = y_{FBF}(n) - y_{MIC}(n) \tag{2}$$

$$y_{FBF}(n) = \frac{1}{M}\sum_{i=1}^{M} x_i(n) \tag{3}$$

식(1)은 GSC 빔포밍 전체의 필터 계수를 구하기 위한 cost function이다. $y_{FBF}(n)$는 식[3]과 같이 time delay값으로 steering된 값들의 합으로 표현되고, $y_{GSC}(n)$는 식[2]에서처럼 Fixed beamformer의 출력값과 MIC단의 출력값의 차로 나타낼 수 있다.

$y_{MIC}(n)$ 의 값은 다음 식[4-6]로부터 얻을 수 있다. M은 입력 microphone의 개수, K는 필터개수, μ 는 learning rate, $u_i(n)$ 는 BM단의 출력값으로 reference를 1번 마이크로폰으로 했을 때의 값으로 식[4]에 표현되어 있다. 여기서, $u_i(n)$는 잡음 기준 신호가 된다. 식[5]의 $\mathbf{n}_l(n)$는 필터계수로서 이 값은 $y_{MIC}(n)$와 $u_i(n)$의 곱을 식[5]와 같이 채널 별로 합하여 학습된 $y_{MIC}(n)$를 구할 수 있다. 필터 계수 $\mathbf{n}_l(n)$의 업데이트는 비용함수 식(!)에 대한 해를 구하기 위한 식(6)과 같이 normalized LMS 알고리듬 방법을 이용할 수 있다. 이렇게 하여 최종적으로 식(2)에 의해서 $y_{GSC}(n)$값을 얻을 수 있게 된다.

$$u_i(n) = x_i(n) - x_1(n), \quad i = 2,...M \tag{4}$$

$$y_{MIC}(n) = \sum_{l=-K}^{K} \left[\mathbf{n}_l(n)\right]^T \mathbf{u}(n-l) \tag{5}$$

$$\mathbf{n}_l(n+1) = \mathbf{n}_l(n) + \mu y_{GSC}(n)\mathbf{u}(n-l) \tag{6}$$

GSC에서는 앞서 설명한 바와 같이 MIC 단계에서 필터를 사용하여 FBF의 출력에 남아 있는 잡음을 부가적으로 제거하게 된다. 그런데, 모든 신호에 대해서 $\mathbf{n}_l(n)$를 업데이트하게 되면 목적 신호의 왜곡이 생길 수도 있기 때문에 목적 신호 부분을 제외시키고 업데이트 시키는 알고리즘을 사용한다. 이를 Adaptation mode controller (AMC)라 부른다.

GSC 빔포밍 알고리즘은 식(3)에서와 같이 Fixed beamforming 구간에서 steering하기 위한 time delay의 정확한 예측이 큰 영향을 미친다. Time-delay가 잘못 예측되면 식(3)에서의 fixed beamforming의 출력 값이 잘못되며 이는 처음부터 잘못된 측정에 따르게 되기 때문이다. 따라서, 본 논문에서는 입력 신호간의 전달함수 차이, 마이크로폰 특성 차이에서 비롯되는 입력신호의 spectrum 차이, 정확한 time delay를 구하는 방법이 가장 중요하다. 이를 위해, 본 논문에서는 Acoustic Transfer Function (ATF)[2-3] 방법을 이용하였다. ATF 방법은 잡음이 없는 환경에서 마이크로폰배열로부터 음성을 녹음하고 이를 식(7)과 식(8)을 이용하여 transfer function의 parameter $\alpha_i(k)$ 를 미리 획득하는 방법이다.

$$J_i(k) = \sum_{n=0}^{T-1} \left| X_r(n,k) - \alpha_i(k) \cdot X_i(n,k) \right|^2 \tag{7}$$

$$\alpha_i^{(opt)}(k) = \frac{\sum_{n=0}^{T-1} \left| X_r(n,k) \cdot X_i^*(n,k) \right|}{\sum_{n=0}^{T-1} \left| X_i(n,k) \right|^2} \tag{8}$$

그림 1 얼굴검출 학습을 위해 사용된 face 영상 예.

2.2 얼굴검출

그림 2 얼굴검출 학습을 위해 사용된 non-face 영상 예.

얼굴검출에 관한 많은 연구가 행해져 왔다[4-7]. Neural Network를 이용한 얼굴 검출[4], Bayesian Discriminating을 이용한 얼굴 검출[5], Support Vector Machine (SVM)을 이용한 얼굴검출[6], hybrid boosting을 이용한 얼굴검출[7]등이 다양한 접근 방법이 있었다. 본 논문에서 사용한 얼굴검출 알고리즘은 조명에 강인한 것으로 알려진 Modified Census Transform (MCT)를 특징으로 하고 Adaboost로 face와 nonface를 학습시켜 얼굴 검출기를 만들었다. 훈련을 위한 DB는 포항공대에서 인터넷 등에서 임의로 수집한 17,000장의 face image와 5,000장의 non-face image 로 구성되어 있다.

3. Experimental Results

본 논문에서는 두 가지 실험을 하였다. 첫번째 실험은 음성인식을 발성하는 목적 음성 방향과 잡음 음성 방향이 서로 다를 경우, 목적 음성의 방향을 얼굴 검출을 통해 획득하고 목적 신호방향으로 빔포밍을 한 후 음성인식을 수행하는 실험이다.
두번째 실험은 다양한 방향의 음원 소스로부터 다양한 방향의 음성인식을 수행할 수 있는 것을 목적으로 하고 있다. 이 논문은 우선 양방향에서 음성 인식이 가능한 단어가 들어오도록 하였다.

그림 4 잡음과 음성이 동시에 모바일 플랫폼(로봇)에 장착되어 있는 마이크로폰에 입력된다. 이 때, 목적음성과 잡음은 서로 다른 각도에서 입력된다.

3-1 잡음환경에서의 음성인식

잡음환경에서의 음성인식은 그림 4와 같이 목적음성과 잡음(스피커)이 마이크로폰(로봇 정면얼굴이 0도)을 기준으로 서로 다른 각도에서 입력된다고 가정한다.
음성인식은 먼저 PBW 454단어 30세트 중 28 set로 훈련하고 음성 잡음을 섞어 두 세트로 테스트한 결과를 표 1에 나타내었다. 이 환경에서의 마이크로폰 배열은 그림 4의 플랫폼에 배치하여야 하므로 지름이 30cm인 타원형에 직사각형으로 배치하였다.

표 1 <잡음 환경에서의 음성인식결과>

SNR	GSC	Noisy
-5	51.99	3.54
0	77.21	20.35
5	86.73	51.33
10	92.48	78.32
15	94.47	87.39
20	92.26	89.82

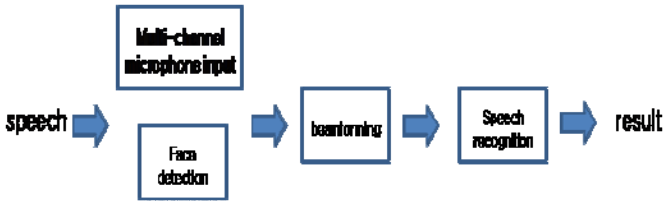

그림 3 빔포밍을 이용한 음성인식 수행 과정

이번에는 잡음을 다양화하여 음성인식을 수행하였다. non-stationary 잡음으로 music은 pop song을, TV소리는 무한도전(한국 코미디 프로그램)으로 대사와 음악과 다양한 잡음이 들리는 소리로 하였다.

표 2 <잡음 환경에서의 음성인식결과>

	Clean	Music	TV
Without BF	100	60	40
BF	100	90	90

목적음성과 잡음의 SNR은 10dB이다. 이 환경에서의 마이크로폰 배열은 그림 5와 같이 마이크로폰을 일렬로 4cm 간격으로 4개를 배치하였다. 본 논문의 모바일 플랫폼에서 자주 사용하는 10단어를 추출하여 음성인식 실험을 한 결과를 표 2에 도시하였다. 빔포밍을 적용한 결과, 잡음 환경에서의 음성인식률이 30%이상 증가함을 확인할 수 있었다.

3-2 양방향 동시 음성인식

두 음원은 마이크로폰을 기준으로 서로 다른 방향에 위치하게 된다. 음원과 마이크로폰의 거리는 3m

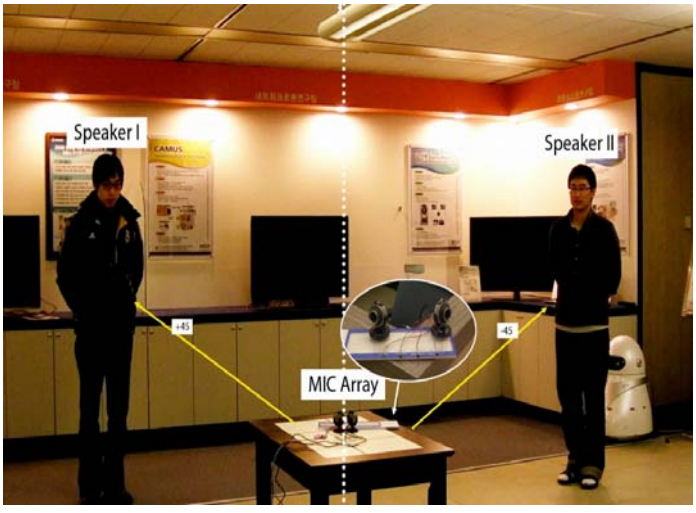

그림 5 두 명의 화자 (speakers)가 마이크로폰을 기준으로 -45도, 45도 각도에 각각 서 있다. 마이크로폰은 저가용 핀 마이크포폰으로 4개를 4cm 간격으로 배열하였다. 실험환경은 일반 office 환경이다.

이하이며 본 실험에서는 2m ~3m 사이에 임의의 위치에 음원이 위치할 수 있다. 본 실험에서는 실험상 편의를 위하여 그림 5와 같이 각각 45 방향과 -45도 방향에 두 화자가 주로 위치하도록 하였으며 화자의 정확한 위치는 얼굴검출을 통하여 획득하도록 하였다. 이 환경에서의 마이크로폰 배열은 3-1의 다양한 잡음환경에서의 마이크로폰 배열과 동일하다.

그림 6 두 사람이 동시에 단어를 발성한다. 두 사람의 방향으로 빔포밍 결과로 음성인식을 수행하였다. 왼쪽 사람은 “상세 일정 알려줘”로, 오른쪽 사람은 “오늘 날씨 어때?”를 동시에 발성하였다.

인식대상 단어의 크기는 3-1의 두 번째와 동일한 10단어를 사용하였다. 두 사람이 동시에 10단어 내의 단어를 발성한다. 이 때, 발성된 한쪽의 발성이 0.5s 빨리 발성할 수도 있고 늦게 끝날 수도 있다. 그리고 두 사람의 사용자는 체감적으로 비슷한 크기의 목소리를 낸다고 가정한다. 예를 들어, 한 사람의 발성의 크기가 120dB이고 다른 사람은 40dB 정도로 큰 차이가 나는 상황은 없다고 가정한다. 이렇게 실험한 결과를 표 3에 나타내었다

표 3 <양방향 동시 음성인식결과>

	양쪽 성공	한쪽만 인식	양쪽 실패
인식률	80	17	3

표 3에서 한쪽만 인식된 경우는 어느 한쪽이 유달리 크게 발성한 경우가 대부분이였다. 비슷한 목소리의 크기로 발성한다는 사실을 두 명의 화자가 서로 알고 있었으나 사람마다 목소리의 성량 차이가 있기 때문에 이와 같은 결과가 나왔다고 보여진다. 양쪽 모두 실패한 경우는 본 논문에서 사용한 음성인식기의 성능 때문으로 사료된다.

4. 결 론

본 논문에서 사용한 방법은 모바일 환경에서 사용자의 방향을 얼굴검출을 통하여 알아내고, 이 방향으로 빔포밍을 하여 성공적으로 잡음을 제거할 수 있음을 보였다. 또한, 두 방향에서 동시에 두 명의 사용자가 음성 명령을 수행하였을 경우에도 얼굴검출을 통하여 양 방향으로 빔포밍을 동시에 수행하고 이를 통해 양방향의 음성을 인식할 수 있음도 보였다.

이를 통해 본 논문의 방법은 모바일, 잡음 환경에서 non-stationary 잡음을 제거하고 음성인식을 성공적으로 할 수 있음을 보였다.

본 논문에서는, 빔포밍에서 중요한 steering을 성공적으로 수행하기 위하여 두 가지 단계를 추가하였다. 먼저, 잡음환경에서 음원추적을 이용하여 음원방향을 알아내기 어렵기 때문에 이를 보완하기 위하여 얼굴검출을 수행하였다. 그리고, 잡음이 없는 환경에서 각 방향에 대하여 정확한 time delay값을 측정하고 마이크로폰 간의 mismatch를 보완해주기

표 2 <잡음 환경에서의 음성인식결과>

	Clean	Music	TV
Without BF	100	60	40
BF	100	90	90

위한 방법으로 ATF를 수행하였다.

ATF 방법은 사전 정보를 사용하여 음성인식률을 10dB 환경에서 14%이상 향상시킬 수 있었으나, 사전에 각도에 따라 학습해야 한다는 단점이 있다. 또 한, 전시장과 같은 0dB 이하의 잡음 환경에서는 빔포밍 결과가 목적음성을 왜곡시킨다. 따라서, Future work으로서 AMC를 수정하여 0dB이하의 잡음환경에서 목적음성을 왜곡시키지 않도록 수정할 예정이며, 사전에 여러 방향에 대해 학습이 필요한 ATF를 사전학습이 필요없는 adaptation이 가능하면서도 수행 성능을 저하시키지 않는 방법을 모색할 예정이다.

5. 참고문헌

[1] O.L Frost, “An Algorithm for Linearly Constrained Adaptive Array Processing,” *IEEE proceedings*, Vol. 60, No. 8, August 1972.

[2] G. Reuven, S. Gannot, and I. Cohen, “Multichannel Acoustic Echo Cancellation and Noise Reduction in Reverberant Environments Using the transfer-function GSC,” *IEEE*, 2007

[3] Israel Cohen, Sharon Gannot, Baruch Berdugo, “An Intergrated Real-Time Beamfoming and Postfiltering System for Nonstationary Noise Envirionments,” EURASIP Journal on Applied

Signal Processing vol.2003 no.11 pp1064-1073 2003
[4] HA Rowley, S Baluja, T Kanade, “Neural network-based face detection,” IEEE trans. on PAMI vol.20 no.1 pp23-38 1998
[5] Chengjun Liu, “A Bayesian Discriminating Features Method for Face Detection,” IEEE on PAMI vol. 25, no. 6 pp.725-740 2003
[6] Yongmin Li , Shaogang Gong , Jamie Sherrah , Heather Liddell, “Support vector machine based multi-view face detection and recognition,” Image and Vision Computing vol. 22 pp.413-427 2004
[7] Hsiuao-Ying Chen, Chung-Lin Huang, Chih-Ming Fu, “Hybrid-boost learning for multi-pose face detection and facial expression,” Pattern Recognition vol.41 no.3 pp1173-1185 2008