음성인식 시스템에서의 원격 음성입력기의 성능평가

이광석
진주산업대학교

A Performance of a Remote Speech Input Unit in Speech Recognition System

Gwang-seok Lee
Jinju National University
email : kslee@jinju.ac.kr

요 약

본 연구에서는, 음성인식 시스템에서의 마이크 어레이 기반으로 한 beamforming 방법을 기반으로 음성신호에 대한 에러감소 알고리듬의 성능평가를 위한 시뮬레이션 하였으며 그 성능을 분석하였다. 또한, 마이크 어레이로부터 취득한 음성신호로 부터 각 채널에 대한 최대 신호대잡음비 구하고 음성신호별로 신호대잡음비를 비교 검토하였다. 음성 인식률은 경우1에서는 54.2%에서 61.4%로, 경우2에서는 더 낮은 신호대잡음비로 41.2%에서 50.5%로 각각 개선됨을 알 수 있었다. 따라서 평균 에러감소율은 경우1에서 15.7%를 보였다.

ABSTRACT

In this research, We simulated performances of error reduction algorithm for the speech signal based on the microphone array-based beamforming method in speech recognition system and analyzed its performance. Also, we processed speech signal adopted from microphone array and maximum signal to noise ratio for each channel, and then compared them with signal to noise ratio of speech signal. Speech recognition rate is improved from 54.2% to 61.4% in case 1 and is improved from 41.2% to 50.5% in case 2 of the lower signal to noise ratio. Therefore the average reduction rates are showed 15.7% in case 1.

키워드

Remote Speech Input, Microphone Array, Automatic Speech Detection

Ⅰ. 서 론

높은 S/N비를 유지해야 하는 음성인식시스템의 경우는 마이크를 가까이에 두고 발성하여야만 하는 문제점이 존재한다. 또한 대부분의 음성인식시스템에서 마이크를 이용하여 음성을 입력하는 방식을 채택하고 있다. 그러나 하나의 마이크로 음성을 입력받는 경우에는 음성입력 시에 마이크의 위치에 항상 세심한 주의를 기울여야 한다. 따라서 일반인들이 실용적으로 사용하기에는 여러 가지로 불편하므로 원격음성입력기는 대화체 음성 번역 시스템이 가지는 이와 같은 음성입력의 문제점을 해결하기 위하여 시도되었다. 즉, 마이크와 화자가 어느 정도의 거리를 두고, 마이크의 위치에 주의를 기울이지 않고 발성할 경우에도 허용된 수준 이상의 신호 대 잡음비를 갖는 음성을 자동으로 입력하는 기능을 한다. 이를 만족시키기 위하여 8채널 마이크 어레이 기반의 원격음성입력기를 구성하여 적용하고 성능을 평가하였다.

Ⅱ. 마이크 어레이와 화자의 상대적 위치에 따른 음성 인식률의 개선

2.1. 평가방법에 따른 음성 인식률 평가-방법1

잡음이 존재하는 실험실에서 남성화자 2명이 발성한 음성원을 사용하였다. 평가방법으로는 8개의 마이크로 구성된 마이크 어레이로부터 40cm 와 80cm 떨어진 위치에서 마이크 어레이의 중앙, 왼쪽 마이크에서 정면 방향, 오른쪽 마이크에서 정면 방향 등 세 곳에서 5개의 한국어 문장을 반복적으로 발성하여 각 채널의 음성 데이터와 이들을 지연 및 가산한 8 채널 마이크 어레이 출력 음성 데이터를 수집하였고 이들로부터 S/N비를 계산하였다. S/N비를 계산하기 위해서 수집된 음성 데이터를 수동으로 음성부와 묵음부를 검출하였다. 또한, 일관성 있는 S/N비를 계산하기 위하여 문장의 발성 속도와 발성 크기에 주의를 기울여 일정하게 발성하였으며 음성부 부분에서는 묵음 부분이 생기지 않도록 연속적으로 발성하였다. 각각의 거리와 위치에서 수집된 5문장의 음성 데이터에서 구한 S/N비로부터 각 채널로부터 수집된 음성 데이터와 8채널 마이크 어레이 출력을 신호 처리한 음성 데이터의 S/N비의 차이를 비교분석하였다. 아래의 결과 항목 중에서 첫 번째 항목은 8채널 중에서 최대 S/N비를 나타내는 한 채널의 5문장에 대한 평균 S/N비를, 두 번째 항은 8채널 마이크 어레이 출력을 지연-가산(Delay and Sum BeamForming :DSBF)신호 처리한 음성의 평균 S/N비를 나타낸 것이다. 세 번째 항목은 두 번째 항목인 8채널 어레이 음성

데이터의 평균 S/N비와 첫 번째 항목인 8채널 중 최대 S/N비인 한 채널의 평균 S/N비의 차로써 8채널 마이크 어레이에 의한 S/N비의 개선을 나타낸 결과이다. 네 번째 항목은 세 번째 항목과 같은 S/N비의 개선이지만 5문장 중에서 최대의 S/N비 개선을 나타낸 한 문장에 대한 S/N비의개선을 나타낸 결과이다. 평가결과를 분석해 보면 약 40cm 의 거리에서 최대 S/N비를 나타내는 1채널의 음성신호에 비해서 8채널의 마이크 어레이 출력을 지연-가산(DSBF)신호 처리한 음성신호는 약2.4~4.0dB의 평균 S/N비의 향상을 나타내었으며, 단일 문장에 대해서는 최대 5.9dB까지의 S/N비의 개선을 얻었다. 80cm의 거리에서 구한 결과들도 비슷한 양상을 보였는데, 최대 S/N비를 나타내는 1채널의 음성신호에 비해서 8채널의 마이크 어레이 출력을 신호 처리한 음성신호는 평균 S/N비에서 약 2.3~3.8dB정도 개선을, 단일 문장에 대해서는 최대 5.4dB의 S/N비의 향상을 얻었다.

표 1. 발화자의 위치에 따른 S/N비 개선[dB]

Speaker's distance and position[cm]		Max.Ch.	8Ch. Array.	Average gain	Max. gain
40	Mid	15.9	19.9	4.0	5.9
	Left	18.6	21.0	2.4	4.4
	Right	17.8	20.9	3.1	4.7
80	Mid	16.7	19.0	2.3	3.8
	Left	15.1	18.3	3.2	4.9
	Right	17.2	21.0	3.8	5.4

2.2. 평가방법에 따른 음성 인식률 평가-방법2

잡음이 존재하는 실험실에서 실험실 환경에서 남성 화자 1명이 발성한 음성원을 사용하였다. 평가방법으로는 한국어 대화체 50문장을 8채널 마이크 어레이에서 60cm 떨어진 세 곳의 다른 위치에서 반복 발성하여 수집한 음성 데이터를 본 연구실에서 개발한 음성인식 시스템으로 평가하였다. 발성 화자는 마이크 어레이의 중심, 왼쪽, 오른쪽으로부터 각각 60cm 거리를 유지한 다음, 동일한 50문장을 반복 발성하였다. 이 실험에 사용된 음성인식 시스템은 본 연구실에서 개발된 대화체 음성번역 시스템에 통합된 음성 인식기 모듈로서, 2000단어의 대화체 화자독립 환경에서 약 72%의 단어 인식률을 나타내고 있다.

표 2. 마이크 어레이와 화자의 상대적 위치에 따른 음성 인식률[%]

Mic.-array의 위치	Left	Mid	Right
8Ch. Mic.-array에 의한 음성 인식률	66.8	72.7	72.6

화자와 마이크 어레이 간의 상대적인 위치에 따른 음성 인식률의 변화를 보면 마이크 어레이의 중앙, 오른쪽에서의 인식률은 서로 비슷함을 알 수 있다. 그러나 왼쪽에서의 음성 인식률은 중앙과 오른쪽에 비해서 차이를 나타냄을 알 수 있는데 이는 마이크 어레이의 왼쪽부분을 이루는 마이크들의 감도의 저하를 들 수 있으며, 또한 각 위치에서 발성된 음성 데이터가 서로 완전히 동일하지 않다는 점에서 기인하였다고 추측된다.

III. S/N비가 상이한 환경에서의 8채널 마이크 어레이의 S/N비 및 음성 인식률 개선

사전에 수집된 백색잡음(TV의 비 방송 채널에서 발생되는 잡음을 마이크를 사용하여 입력한 후 DAT에 저장) 데이터를 스피커를 통해 잡음의 크기를 달리 발생시켜 S/N비가 상이한 두 가지의 실험실 환경(Case1, 2)을 조성하였고 각각의 환경에서 남성 회자 5명이 한국어 대화체 50문장을 각각 발성하여 수집된 두 종류의 음성 데이터를 실험에 사용하였다. 이때 화자의 마이크 어레이 간의 거리는 약 60cm 정도가 되도록 조정하였다. 각 화자가 발성한 음성 데이터들로부터 S/N비 및 음성 인식률의 개선정도를 조사하였다. 먼저 S/N비의 개선정도를 화자별로 조사하였다. 또한 상이한 S/N비 환경에서 마이크 어레이에 의한 음성 인식률의 개선도를 각 화자별로 조사하였다. 표3은 두 가지 상이한 S/N비 상황에 대해서, 8채널 마이크 중에서 무작위로 5문장을 선택하여 구한 S/N비 값들이다. 먼저 Case 1은 1채널의 평균 S/N비가 21.8dB인 환경으로서, 8채널 마이크 어레이를 이용한 지연-가산 신호처리에 의해 평균 4.3dB의 S/N비의 개선을 얻었다. 신호대음비가 상대적으로 낮은 Case 2에서는 8채널 마이크 어레이를 이용한 지연-가산 신호처리에 의해 약 2.2dB의 개선을 얻었다.

표4에 나타난 결과는 표3에서와 같이 상이한 두 가지 S/N비 환경에 대해 8채널 마이크 어레이에 의한 음성 인식률 개선 정도를 조사한 실험의 결과이다. 결과를 보면 각 화자별로 음성 인식률에서 약간의 편차를 나타내고 있다. 5명의 화자를 대상으로 한 실험의 결과를 보면 1채널 음성신호의 평균 S/N비가 21.8dB인 Case 1에서는 음성 인식률이 54.2%에서 61.4%로 개선되어 약 15.7%의 평균 오류 감축률을 나타내었으며 이보다 낮은 S/N비 환경인 Case 2에서의 음성 인식률은 41.2%에서 50.5%로서 전체적인 음성 인식률에서는 약간의 저하를 나타내었지만 평균 오류 감축률에서는 약 16.0%로 Case 1과 비슷한 수치를 보였다. 따라서 마이크 어레이를 이용하면 음성신호를 개선하면 주변 S/N비에 구애받지 않고 음성 인식률에서 일정한 수준의 오류 감축 효과를 얻을 수 있음을 알 수 있었다.

IV. Close-talk 방식의 단일 마이크와 마이크 어레이 간의 음성 인식률 차이 비교

본 실험에서는 마이크와 화자의 거리가 20cm 이내를 유지하는 close-talk 방식의 단일 마이크 음성 입력에 의한 음성 인식률 및 마이크 어레이가 기반의 원격음성입력방식에서의 마이크 수에 따른 음성 인식률을 구하여 각각의 방식에서의 음성 인식률 차이를 조사하였다. 본 실험을 위하여 남성화자 2명이 각각 발성한 한국어 대화체 50문장을 실험실 환경에서 수집하여 음성 데이터로 사용하였다.

표 3. 두 가지 S/N비 환경에 대한 8채널 마이크 어레이의 S/N비의 개선[dB]

Speaker	Case 1			Case 2		
	Max. SNR Ch	8 Ch DSBF	SNR gain	Max. SNR Ch	8 Ch DSBF	SNR gain
KBC	20.6	26.0	5.4	12.9	17.2	4.3
	23.8	26.8	3.0	13.3	15.2	1.9
	21.8	25.7	3.9	15.1	17.1	2.1
	19.8	24.8	5.0	14.3	15.1	0.8
	19.9	25.3	5.4	13.7	18.9	5.3
LGS	25.3	28.1	2.8	14.6	18.2	3.6
	25.3	29.7	4.4	12.5	18.6	6.1
	26.2	29.2	3.0	17.2	19.8	2.6
	27.2	30.5	3.3	15.5	18.7	3.1
	25.2	30.2	5.0	15.4	17.5	2.0
KKP	19.7	22.8	3.1	16.2	17.9	1.7
	18.8	22.3	3.6	16.3	17.5	1.2
	18.6	22.6	4.0	9.0	12.3	3.2
	20.5	23.8	3.3	13.5	14.1	0.5
	20.0	23.6	3.6	14.5	15.0	0.5
SWW	21.0	23.2	2.3	10.0	13.0	2.9
	22.2	26.6	4.4	12.5	14.1	1.6
	20.8	24.2	3.4	12.5	15.1	2.7
	21.6	25.3	3.7	14.0	15.0	1.0
	22.5	27.0	4.6	12.5	14.1	1.7
KKS	19.6	28.4	8.8	15.2	16.2	1.1
	19.0	26.9	7.9	14.5	17.1	2.3
	20.3	24.4	4.1	15.2	15.6	0.4
	23.9	28.9	5.0	18.4	19.2	0.8
	22.6	26.2	3.6	12.3	14.0	1.7
Average	21.8	26.1	4.3	14.0	16.3	2.2
Standard dev.	2.5	2.4	1.5	2.1	2.1	1.5
Maximum	27.2	30.5	8.8	18.4	19.8	6.1
Minimum	18.6	22.3	2.3	9.0	12.3	0.4

표 4. 두 가지 S/N비 환경에 대한 8채널 마이크 어레이의 음성 인식률 개선[%]

	Case 1			Case 2		
Speaker	Max. SNR Ch	8 Ch DSBF	ERR	Max. SNR Ch	8 Ch DSBF	ERR
KBC	35.7	46.6	17.0	19.6	26.9	9.1
LGS	57.6	63.5	13.9	49.6	62.7	26.0
KKP	60.6	64.9	10.9	52.2	54.2	4.2
SWW	64.2	70.6	17.9	55.4	68.1	28.5
KKS	52.8	61.5	18.4	28.5	40.4	16.6
Average	54.2	61.4	15.7	41.1	50.5	16.0

▪ ERR(Error Reduction Rate) = (e1-e2/e1)×100

본 실험의 평가방법은 기존의 8개의 마이크로 구성된 마이크 어레이에서 하나의 마이크를 선정하여 별도로 화자와 가까운 거리를 유지하도록 마이크 스탠드에 고정시켜 close-talk용으로 사용하였고 나머지 7개의 마이크로 이루어진 마이크 어레이를 구성하고 마이크 어레이의 중심에서 약 60cm 정도, close-talk 마이크와 20cm 이내의 거리를 유지하면서 발성하게 하였다. 이와 같이 발성된 음성에 대하여 close-talk 마이크 신호, 마이크 어레이에서 2, 4, 7채널의 지연-기산 신호 처리한 신호를 각각 수집하여 음성 인식률의 차이를 구하였다. 마이크 어레이의 2, 4, 7채널의 선정은 S/N비가 높은 순서로 정하였다.

표5. Close-talk 단일 마이크 입력방식과 채널 마이크 어레이를 이용한 원격음성입력 방식에서의 음성 인식률 차이

Speaker	1Ch. Close-talk		Mic.-array			
			2Ch.	4Ch.	7Ch	
	R.R[%]	SNR[dB]	R.R[%]	R.R[%]	R.R[%]	R.R[%]
SWW	67.6	24.8	60.0	62.9	65.6	21.8
KKP	74.7	19.4	60.6	61.7	67.7	17.7
Average	71.2	22.2	60.3	62.3	66.7	19.8

표4는 Close-talk 방식의 단일 마이크를 사용한 경우와 마이크 어레이를 이용한 원격음성입력방식의 경우에 대한 음성 인식률의 변화를 조사한 결과이다. 본 실험의 결과를 보면 60cm 거리에서 7채널의 마이크 어레이를 이용한 원격음성 입력 방식이 약 20cm의 close-talk 음성 입력방식에 비해 낮은 음성 인식률을 보임을 알 수 있다. 채널수에 따른 마이크 어레이의 음성 인식률은 채널수가 증가할수록 인식률의 개선정도가 커짐을 알 수 있다. 비록 높은 S/N비의 음성 데이터라고 하더라도 화자에 따라서는 낮은 음성 인식률을 나타냄을 알 수 있다.

이와 같은 변화는 화자 간의 발성 특성에서 기인한다고 추측된다. 본 실험에서는 음성입력 하드웨어의 채널수 제한으로 인하여 close-talk 방식에 의한 음성입력과 8채널 마이크 어레이를 이용한 원격음성 입력간의 음성 인식률 차이를 직접 조사하지는 못했다. 그러나 채널수의 증가에 따른 마이크 어레이의 인식률의 증가를 고려하면, 8채널 마이크 어레이에 의한 원격음성입력방식은 7채널에 의한 방식에 비해 close-talk 단일 마이크 입력 방식과의 인식률의 차이를 더욱 줄일 수 있을 것으로 예상된다.

V. 결 론

본 연구에서는 8채널의 마이크 어레이를 이용한 음성의 잡음감소 알고리듬의 성능을 여러 평가방법으로 평가하고 그 성능을 분석하였다. 각각의 채널별 음성의 최대 S/N비와 마이크 어레이로부터 얻어진 음성신호를 신호 처리하여 구한 음성의 S/N비를 구한 후 비교분석하였으며 마이크 어레이와 close-talk 방식의 단일 마이크 음성입력에 의한 음성 인식률의 차이를 비교분석하였다.

평가결과를 살펴보면 약 40cm 의 거리에서 최대 S/N 비를 나타내는 1채널의 음성신호에 비해서 8채널의 마이크 어레이 출력을 지연-가산 신호 처리한 음성신호는 약2.4~4.0dB의 평균 S/N비의 향상을 나타내었으며 단일 문장에 대해서는 최대 5.9dB까지의 S/N비 개선을 얻었을 수 있었다. 또한 화자와 마이크 어레이 간의 상대적인 위치에 따른 음성 인식률의 변화를 보면 마이크 어레이의 중앙, 오른쪽에서의 인식률은 서로 비슷함을 알 수 있었다. 표3과 표4의 실험결과를 보면 60cm 거리에서 7채널의 마이크 어레이를 이용한 원격음성 입력 방식이 약 20cm의 close-talk 음성 입력방식에 비해 낮은 음성 인식률을 보임을 알 수 있다. 채널수에 따른 마이크 어레이의 음성 인식률은 채널수가 증가할수록 인식률의 개선 정도가 커짐을 알 수 있다. 비록 높은 S/N비의 음성 데이터라고 하더라도 화자에 따라서는 낮은 음성 인식률을 나타냄을 알 수 있다. 이와 같은 변화는 화자 간의 발성 특성에서 기인한다고 추측된다.

한편 본 연구에서는 음성 입력 하드웨어의 채널 수 제한으로 인하여 close-talk 방식에 의한 음성입력과 8채널 마이크 어레이를 이용한 원격음성입력간의 음성 인식률 차이를 직접 조사하지는 못했다. 그러나 채널수의 증가에 따른 마이크 어레이의 인식률의 증가를 고려하면 8채널 마이크 어레이에 의한 원격음성입력방식은 close-talk 단일 마이크 입력 방식과의 인식률의 차이를 더욱 줄일 수 있을 것으로 예상된다.

참고문헌

[1] J.-W. Yang and Y. Lee, 1996, "Toward Translation Korean Speech into Other Languages," *Proc. ICSLP 96*, vol.4, pp.2368-2370, Oct.

[2] S.U.Pillai, 1989, *Array Signal Processing*, pring-Verlag, New York, pp.8-18.

[3] R.P. Ramachandran and R. J. Mammone, 1995, "Microphone Array for Hands-free Voice Communication in a Car,"*Modern Methods of Speech Processing*, KAP, Boston.

[4] J.L. Flanagan, L.D. Johnstone, R. Zahn, and G.W. Elko, 1985, "Computer-Steered Microphone Arrays for Sound Transduction in Large Room," *Journal Acousitical Society of America*, vol.78, pp.1508-1518, Nov.

[5] D. Giuliani, M. Matassoni,M. Omologo and .Svaizer, 1995, "Roubust Continuous Speech Recognition using a Microphone Array," *Proc. EUROSPEECH*, vol.3,pp.2021-2024.

[6] H. Lee and M.Hahn, 1993, "Development of a Real-time Endpoint Detection Algorithm," *Proc. ICSPAT*, vol.2, pp. 1547-1553, Sep.

[7] G.C. Carter, 1987, "Coherence and Time Delay Estimation", *Proceedings of the IEEE*, vol.75, p.236-255, Feb.