강인음성인식에서 beamformit 알고리즘의 각 단계별 영향 분석

Analysis of the steps included beamformit algorithm in robust speech recognition

권 해 용, 박 형 민
서강대학교 전자공학과

haeyongk@sogang.ac.kr hpark@sogang.ac.kr

Abstract

CHiME 챌린지 등에서 baseline에 사용되는 전처리 기법인 Beamformit 알고리즘[1]은 reference 채널 선택을 위한 average cross correlation 활용, noise threshold를 이용한 silence/noise 필터링, 모든 채널에 대해 수행되는 2단계 Viterbi decoding, 잡음 필터와 local correlation을 반영한 weighted channel summation 등 여러 단계로 구성되어 있다. 우리는 이러한 Beamformit의 구성요소 중 강인음성인식에 주로 기여하는 부분이 어느 것인지 알아보고자 한다. 중점적으로, reference 채널 선택 알고리즘의 유무, noise threshold 결정 방식의 차이, 2단계 Viterbi decoding 알고리즘의 유무가 CHiME 4 challenge[2]에서 사용하는 음성 데이터셋의 음성인식률에 미치는 영향을 비교, 평가하였다. 결과적으로, 음성인식률에 큰 영향을 끼치는 모듈은 reference 채널 선택 알고리즘인 것을 확인하였다.

Motivation

- Beamformit[1]은 CHiME 챌린지에서 강인한 인식 성능을 낸다고 알려짐[2]
- Beamformit의 어떤 구성요소가 음성인식률 향상에 주로 기여하는지에 대해서는 분석된 바가 거의 없음
- Beamformit 논문[1]에서는 주로 speaker diarization에 대해 다루고 있고, speech recognition 성능에 대해서는 단순히 결과만을 언급
- Beamformit의 구성요소 중 어떤 것이 CHiME 4 task에 대한 강인한 인식성능에 기여하는지 파악하고자 함

An outline of Beamformit

- Average cross correlation을 이용한 reference 채널 선택

약 1초 길이의 segments로 windowing ⇨ 각 채널에 대해 다른 채널과의 average GCC를 계산 ⇨ 가장 큰 average GCC 값을 가지는 채널을 reference 채널로 선택

- Silence/Noise filtering with relative noise threshold

약 0.5초 길이의 segments로 windowing ⇨ Reference-다른 채널 pair에 대해 프레임 별 GCC를 계산 ⇨ 모든 프레임의 GCC를 정렬했을 때 (상위 10%의 값)/(마이크개수-1)을 threshold로 설정

Threshold보다 작은 GCC를 가지면 noise

Threshold보다 작은 GCC를 가지지만 앞에 위치한 프레임의 경우 silence

- N-best tdoa extraction

약 0.5초 길이의 segments로 windowing ⇨ 채널 pair에 대해 프레임 별 GCC를 계산 ⇨ 프레임 별 GCC 에서 서로 근접하지 않는 N개의 최대 peak 값을 추출

- 2-step Viterbi decoding

프레임 별 GCC를 emission score로, 다음 프레임과의 tdoa 차이를 transition score로 설정 ⇨ 채널 pair 별 독립적으로 Viterbi decoding 수행, 2-best tdoa path 추출 ⇨ 채널 pair와 2-best states로부터 가능한 조합을 하나의 새로운 state로 고려하여 Viterbi decoding 수행, 최종 tdoa path 추출

- Weighted channel summation

Best tdoa로 보정된 채널별 신호에 대해 local xcorr 계산, 이를 바탕으로 채널 별 weight 계산 ⇨ Weighted channel summation

Experimental setup

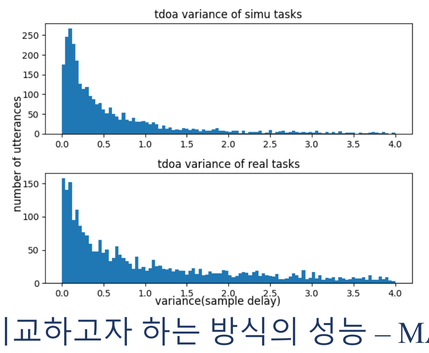
실험 프레임워크: kaldi CHiME 4 baseline with all channels(1, 2, 3, 4, 5, 6)
음성인식 모델: trained GMM-HMM (CHiME 4 official)
알고리즘 소스코드: MATLAB-implemented beamformit[6]
평가지표: CHiME 4 챌린지 각 task에 대한 100-WER

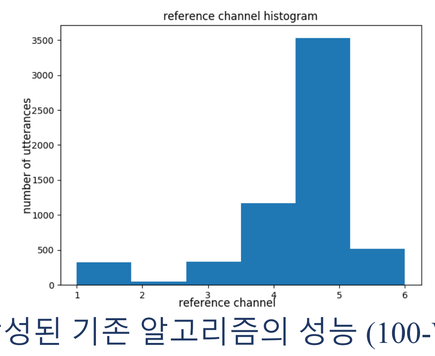
비교 항목은 다음과 같으며, C++ 기반 원본 알고리즘과의 비교를 제외하면 전부 MATLAB 코드를 수정하여 이루어졌다.

- 기존 C++ 기반 알고리즘[5] 과 새로 작성한 MATLAB 기반 알고리즘 인식 성능
- Reference 채널 선택 시 segment 개수 별 인식 성능
- Reference 채널 선택 알고리즘 유무에 따른 인식 성능
- Viterbi, clustering 적용 유무에 따른 인식 성능
- silence, noise filter 적용 유무에 따른 인식 성능
- Weighted channel summation 적용 유무에 따른 인식 성능

Experimental results

- Real/Simulation 별 tdoa 분산의 히스토그램, 선택된 reference 채널의 분포





- 비교하고자 하는 방식의 성능 – MATLAB으로 작성된 기존 알고리즘의 성능 (100-WER)

새로 작성한 MATLAB 기반 알고리즘과 기존 C++ 기반 알고리즘[4]의인식 성능 차이(C++ ver. – MATLAB ver.)

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-0.09	0.14	0.03	-0.07
dt05_real	0.05	-0.18	0.16	0.00
et05_simu	-0.13	-0.38	-0.77	-0.53
et05_real	-0.36	-0.08	0.17	0.56

Reference 채널을 여섯 번째 채널(best case)로 고정하였을 때

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.03	-0.11	-0.12	-0.61
dt05_real	-0.68	-0.11	-0.41	0.40
et05_simu	-0.15	-0.09	0.54	0.13
et05_real	-0.95	-0.78	2.04	-0.07

두번째 채널(worst case) 신호를 그대로 음성인식에 사용한 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-5.47	-17.32	-5.24	-12.05
dt05_real	-52.50	-44.46	-45.92	-63.60
et05_simu	-4.26	-8.57	-7.04	-3.32
et05_real	-57.22	-63.43	-61.77	-60.62

다섯 번째 채널 신호를 그대로 음성인식에 사용한 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-17.05	-21.79	-17.15	-21.51
dt05_real	-9.61	-7.05	-3.88	-5.60
et05_simu	-13.22	-20.15	-17.87	-14.06
et05_real	-20.15	-16.58	-10.84	-8.72

Best tdoa를 그대로 사용하였을 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.09	-0.11	0.10	-0.05
dt05_real	-0.24	0.42	-0.13	0.08
et05_simu	-0.04	0.00	-0.15	0.08
et05_real	0.38	-0.22	0.17	-0.26

Noise filter를 적용하지 않았을 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.15	0.13	-0.01	0.16
dt05_real	0.26	0.49	0.08	0.27
et05_simu	0.47	0.92	1.22	1.03
et05_real	0.15	0.08	0.77	0.04

전체 프레임에 대해 tdoa의 최빈값만을 사용한 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-0.63	-0.34	-0.72	-0.62
dt05_real	-0.59	0.23	-0.26	-0.10
et05_simu	-1.35	-0.24	-1.79	-1.51
et05_real	-0.11	-0.24	0.17	-0.28

Reference 채널 선택 알고리즘에서 사용한 piece 개수가 10일 때

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.03	-0.37	0.03	-0.08
dt05_real	-0.20	0.29	0.11	0.08
et05_simu	-0.41	-0.24	-0.35	0.13
et05_real	0.10	-0.29	-0.11	-0.18

Reference 채널을 두 번째 채널로 고정하였을 때(worst case)

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.17	-0.68	0.01	-1.07
dt05_real	-5.92	-2.46	-1.59	-7.44
et05_simu	0.07	-0.24	-1.49	-0.71
et05_real	-12.97	-7.97	-11.14	-11.03

Reference 채널을 그대로 음성인식에 사용한 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-4.67	-6.80	-5.25	-3.54
dt05_real	-7.53	-5.08	-3.14	-4.07
et05_simu	-2.90	-3.67	-2.20	-1.96
et05_real	-15.70	-13.14	-7.83	-5.67

네 번째 채널 신호를 그대로 음성인식에 사용한 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-7.33	-5.18	-4.17	-3.26
dt05_real	-9.64	-4.18	-2.83	-3.60
et05_simu	-5.19	-4.12	-1.44	-0.11
et05_real	-16.77	-11.67	-5.40	-4.89

Clustering이 적용되지 않은 1-step Viterbi 만을 사용하였을 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	0.05	0.08	0.06	-0.03
dt05_real	-0.21	-0.04	-0.03	-0.26
et05_simu	-0.06	-0.03	-0.02	0.04
et05_real	0.17	-0.01	-0.37	-0.52

Fixed noise threshold를 사용하였을 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-0.25	-0.18	-0.22	-0.36
dt05_real	-0.98	0.15	-0.39	-0.53
et05_simu	-0.54	-0.99	-0.61	-0.37
et05_real	-0.18	-0.44	0.11	-0.16

Weighted channel sum을 하지 않았을 경우

	BUS	CAFE	PEDESTRIAN	STREET
dt05_simu	-0.11	-0.15	-0.09	-0.05
dt05_real	-0.21	-0.08	0.05	0.11
et05_simu	-0.34	-0.61	-0.65	-0.46
et05_real	-0.80	0.02	-1.85	-0.58

Conclusion & Discussion

- **음성인식률 향상에 가장 큰 영향을 끼치는 것은 reference 채널 선택 알고리즘**
- **Reference 채널을 잘못 선택했다 하더라도 다른 요소에 의해 다소 보완**
- 각 요소 별로 인식 성능 향상에 조금씩 기여하는 것을 확인
- Real 환경에서 추정된 tdoa의 분산이 simulation에 비해 크다는 것을 확인
- 2-step viterbi decoding의 영향을 확인하기 위해 CHiME5 DB에 대한 실험 필요

References

[1] Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing, volume 15, number 7, pp.* 2011-2023.

[2] Vincent, E., Watanabe, S., Nugraha, A., Barker, J., & Marxer, Ricard. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech and Language, volume 46, pp.* 535-557.

[3] Anguera, X., Woofers, C., & Hernando, J. (2005). Speaker diarization for multi-party meetings using acoustic fusion. *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop, pp.* 426-431.

[4] Watanabe, S., Delcroix, M., Metze, F., & Hershey, J. R. (2017). New Era for Robust Speech Recognition.

[5] https://github.com/xanguera/BeamformIt

[6] https://github.com/gogyzzz/beamformit_matlab