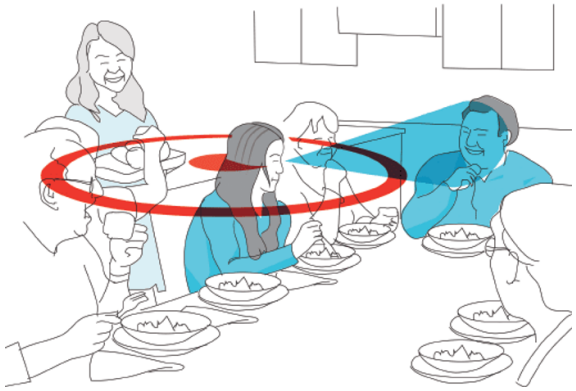
디지털휴먼엔터테인먼트특론 과제 2

120220210 고재현

2023년 3월 15일

1 개요

디지털 엔터테인먼트 분야에서, VR 및 AR 기술은 중요한 산업 분야이다. 디스플레이 기술이 발전함에 따라, 해당 기술에서의 시각적 사용자 경험은 중대되고 있으나, 청각적 사용자 경험을 고려하지 않는다면, 사용자 경험의 완성도는 떨어질 수 밖에 없다. 예를 들어 VRchat과 같이 여러 사용자가 함께 대회를 나누는 경우, 칵테일 파티 문제가 발생할 수 있다. 칵테일 파티 문제는 시끄러운 환경에서 여러 명의 사람들이 대화를 나누는 상황에서, 자신이 듣고자 하는 말소리를 구분하기 어려워지는 현상을 의미한다. 이러한 문제를 해결하기 위해서는, 사용자가 원하는 방향의 소리만을 강화하는 기술이 필요하다. 이러한 기술을 **Directional Hearing** 라고 하며, 본 논문에서는 이러한 기술이 VR/AR 환경에서 적용되도록 경량화한 모델을 제안하려고 한다. directional hearing의 상황은 그림 **??**과



〈그림 1〉 Directional Hearing Case

같이 표현할 수 있다.

2 latency requirements

[?]에서는 두 명의 화자가 대화를 나누되, 헤드셋을 통하여 소리를 전달하며 임의의 지연 시간을 추가하는 방법으로, 청력 기관이 허용하는 지연 시간의 한계를 추정하였다. 그 길이는 대략 20ms이다. [?] 에서도 AR 기기를 이용하여 유사한 실험을 진행하였는데, 같은 결과를 보였다.

청각적 사용자 경험뿐만 아니라, VR/AR 기기의 기능적 한계로 인해 20ms라는 요구사항을 맞추기 위해서는

- 계산에 필요한 리소스는 줄이면서, Latency도 20ms 이내여야 하고
- 계산 능력이 떨어지는 단일 소스기기에서 처리 되어야 함
- 입력 소스는 Array mic 신호 및 각도 정보이며, 바라보는 방향의 소리를 분리 및 향상해야 한다.

3 기존 연구

지난 몇 년간, 이러한 문제를 해결하기 위한 다양한 연구들은 주로 Beamforming을 기반으로 진행되었다. [?, ?, ?]. 빔포밍과 같이 신호 처리를 기반으로 한 기술은 계산량이 적지만 성능이 원하는 만큼 높지 않다는 단점이 있다. 최근 연구들은 [?, ?] 딥러닝 기반의 기술들을 제안하여 Speech Separation 분야에서 뛰어난 성능을 보여주었다. 예를 들어 [?, convTasNet]에서는 TCN, LN, MobileNet 등의 테크닉들을 활용하여 각 Speech Source에 대한 Mask를 추정하는 방식으로 Speech Separation 분야에서 뛰어난 성능을 보여주었다.

하지만, 딥러닝 기반의 기술들은 계산량이 많아 실시간 처리가 어렵거나, 모델의 크기가 커서 모바일 기기에서의 사용이 어렵다는 단점이 있다.

3.1 HybridBeam

이 분야에서의 SOTA 모델은 AAAI 2022에 발표된 [**?**, HybridBeam]이다. HybridBeam은 빔포밍 기반의 전처리를 통해 딥러닝 모델의 연산량을 낮추는 방법으로 모델의 성능을 개선하였다.

4 제안 방법