인공지능 스피커의 반응 속도 개선을 위한 음성 압축 가속기의 구현

*이훈종[1], 유준환[1], 구건재[2]
[1]홍익대학교 전자전기공학부, [2]고려대학교 컴퓨터학과
e-mail: cli3846@naver.com, ujunhwan@naver.com, gunjaekoo@korea.ac.kr

Audio Compression Accelerator Design for Improving the Response Time of AI Speakers

*Hun-Jong Lee[1], Jun-Hwan Yoo[1], Gunjae Koo[2]
[1]School of Electronic and Electrical Engineering, Hongik University
[2]Department of Computer Science and Engineering, Korea University

Abstract

Response time is one of the critical performance factors of artificial intelligence (AI) speakers. The internet network delays and the processing time on cloud server infrastructure dominate the response delays of AI speakers. The network delay is proportional to the size of packets that include the recorded queries. Normally this recorded sound data is not compressed since compression processes can be a heavy burden for the wimpy processors embedded in AI speakers.

In this work we design an audio compression accelerator which can reduce the packet size of user queries. We implement the proposed accelerator on the FPGA-based SoC development board. Our evaluation reveals that the overall response time of an AI speaker is effectively reduced with the audio compression accelerator.

I. 서론

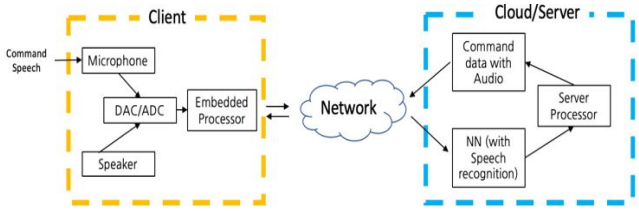
인공지능 스피커에서 반응 시간, 즉 사용자의 질의를 받아서 인공지능 스피커가 답을 제공하기까지의 시간은 인공지능 스피커의 성능을 결정하는 매우 중요한 요소이다. 사용자의 질의 분석 및 음성 처리는 매우 무거운 연산을 요구하기 때문에 클라이언트 단의 인공지능 스피커는 기본적으로 사용자의 음성 신호를 녹음해서 클라우드 서버 단으로 전달하는 역할만을 수행한다. 그러므로, 인공지능 스피커는 대부분 원가 절감을 위해서 저성능/저전력 임베디드 프로세서를 사용하고 있다 [1]. 기본적으로 음성데이터는 압축 효율이 높은 특징을 가지고 있지만 인공지능 스피커는 저성능 프로세서를 사용하고 있기 때문에 음성 데이터를 압축하지 않고 그대로 서버 단으로 전송하는 방식을 사용하고 있다. 이는 서버 단에서의 음성 분석 시간이 상대적으로 오래 걸리기 때문에 서버와 클라이언트 사이의 네트워크 지연시간은 전체 응답 시간에서 그리 중요하지 않다고 여겨 지기 때문이다. 그렇지만, 서버에서의 처리 시간이 단축될수록 이 네트워크 지연시간은 전체 반응 시간에서 무시할 수 없는 시간을 차지하게 된다.

본 논문에서는 인공지능 스피커와 클라우드 서버 사이의 음성 데이터 전송을 위한 네트워크 지연시간을 감소시키기 위한 방법으로 음성 압축 가속기를 구현한다. 인공지능 스피커의 저성능/저전력 임베디드 프로세서로는 압축에 필요한 지연 시간이 길어지게 되므로 이 논문에서는 FPGA를 이용한 가속기를 제안한다. 제안된 가속기는 임베디드 프로세서를 포함하고 있는 FPGA를 이용하여 성능을 분석하였다.

II. 배경

2.1 인공지능 스피커 모델

그림 1은 전형적인 인공지능 스피커의 동작 모델을 나타내고 있다. 인공지능 스피커는 사용자의 음성 데이터를 마이크를 통해서 입력 받아서 이 음성 데이터를 클라우드 서버로 보내는 방식을 가지고 있다. 서버로 보내어진 음성 데이터는 서버 단에서 인공 신경망 (neural network) 알고리즘을 적용하는 음성 분석 과정을 거쳐서 사용자의 질의를 파악하고 질의에 대한 답변 데이터를 인공지능 스피커로 보내게 된다. 사용자의 질의 처리는 매우 강력한 컴퓨팅 연산을 요구하기 때문에 클라이언트 단에 있는 인공지능 스피커는 단지 사용자의 음성을 서버로 전달하는 역할을 수행한다. 음성 데이터는 압축되지 않은 형태로 전달되며 그렇기 때문에 클라이언트와 서버 단 사이의 네트워크 지연에 의해서 많은 시간이 소요되게 된다.



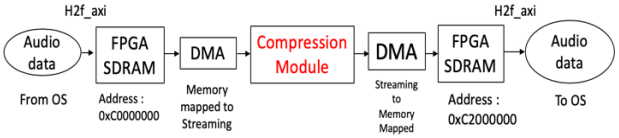
[그림 1. AI 스피커의 동작 구조]

2.2 음성 압축 알고리즘

음성 데이터에 대한 네트워크 지연을 줄이는 효과적인 방법 중의 하나는 음성 데이터를 압축하여 데이터 패킷 사이즈를 줄여서 전송하는 것이다. 음성 데이터 압축 방법 중에서 이 논문에서는 μ-law 알고리즘을 사용하였다. μ-law 방식은 손실 압축 방식 중의 하나로서 알고리즘이 단순하고 구현하기 쉬운 장점이 있다 [2][3]. 본 논문에서는 32-bit의 음성데이터를 16-bit의 코드 (chord)와 스텝 (step)의 형태를 가지는 μ-law 압축 알고리즘으로 음성 압축 알고리즘을 구현하였다.

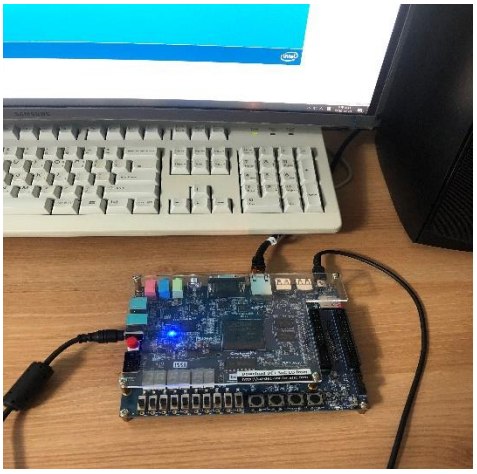
III. 구현

음성 압축 가속기를 구현하기 위해서 본 연구에서는 Intel FPGA의 Cyclone V FPGA SoC가 장착된 DE1-SoC 보드를 사용하였다 [4]. 구현에 사용된 FPGA SoC는 ARM Cortex-A9 임베디드 프로세서를 내장하고 있다. 이는 널리 사용되는 인공지능 스피커인 Google Home Mini에 들어간 프로세서와 동일한 규격이다. 음성 압축 가속기는 FPGA 파트에 구현을 하였으며, 이는 FPGA SoC의 내부에 구현되어 있는 Avalon 인터페이스를 통하여서 임베디드 프로세서가 메모리에 저장한 음성데이터를 받아서 처리한다 [5][6].



[그림 2. 음성 압축 가속기에 대한 데이터 흐름]

그림 2는 FPGA SoC에 구현한 가속기의 구조를 간략하게 나타낸 것이다. 사용자의 음성데이터는 DE1-SoC 보드에 연결되어 있는 마이크와 ADC를 통하여서 외부메모리 (DDR3 SDRAM)에 저장된다. 이 음성 데이터는 데이터 전송 어플리케이션에 의해 FPGA 내부메모리 (FPGA SDRAM)으로 전송된다 [7]. FPGA에 구현한 음성 압축 가속기는 DMA 프로토콜을 이용하여 FPGA 내부메모리의 음성데이터를 읽어서 음성 압축을 수행한다. 압축된 데이터는 역시 DMA 프로토콜을 사용하여 FPGA 내부메모리에 저장되며 데이터 전송 어플리케이션이 이를 읽어서 운영체제에 음성 데이터 압축이 완료되었음을 알려준다 [8]. 이후에 압축된 음성 데이터는 네트워크 전송 어플리케이션을 통하여 서버 단으로 전달되게 된다.



[그림 3, 실험 환경 및 DE1-SoC 보드]

IV. 실험 및 분석

4.1 네트워크 지연시간 분석

음성 압축 가속기의 성능 분석에 앞서 네트워크 지연 시간이 인공지능 스피커의 성능에 미치는 영향을 분석하기 위해서 이 연구에서는 클라이언트와 서버 사이의 데이터 전송 속도를 측정하였다. 음성 데이터를 전송하는 클라이언트로는 서울에 위치한 연구실 내에 설치된 DE1-SoC 보드를 사용하였으며, 음성 데이터를 받는 클라우드 서버로는 미국 Ohio에 위치해 있는 Amazon AWS EC2에 서버 프로그램을

구현하여서 전송 시간을 측정하였다 [10].

음성 데이터의 크기는 80 KB (약 5초 정도의 음성 데이터)라고 가정하였으며 제안된 음성 압축 가속기에 의해 압축이 될 경우 기본적으로 50%의 압축 효율을 가진다. DE1-SoC에서 전송 시간을 측정한 결과 원본 데이터의 경우 692 ms, 압축된 데이터의 경우 534 ms의 시간이 네트워크의 데이터 패킷 전송으로 소요된다. 데이터의 크기가 50%로 줄었지만 네트워크 전송 속도가 23%만 줄은 이유는 인터넷 상에서 동작하는 네트워크 프로토콜의 오버헤드 때문이다.

4.2 음성 압축 가속기의 성능 측정

제안하는 음성 압축 가속기는 Intel FPGA의 Cyclone V FPGA SoC가 장착된 DE1-SoC 보드에 구현하였다. Cyclone V FPGA에 내장된 ARM Cortex-A9 프로세서를 이용하여 Linux 운영체제를 구동하였고, 그 위에 DE1-SoC 보드에 장착되어 있는 마이크 및 ADC를 통하여 음성을 녹음하였다. [8] 제안된 음성 압축 가속기의 성능은 압축 없이 음성을 보내는 설정 (기본 구성)과 FPGA 내부의 두번째 ARM 코어를 사용하는 설정 (소프트웨어 압축)과 성능을 비교하였다. 80 KB의 음성데이터를 압축하는 데에 걸리는 시간은 다음과 같다.

소프트웨어 (Cortex A9)	하드웨어 (가속기)
33.68 ms	17.76 ms

음성 압축을 FPGA에 구현한 하드웨어 가속기로 실행할 경우 소프트웨어적으로 압축할 때보다 1.9배의 성능 향상을 보이고 있다. 실험에 사용한 ARM 코어가 FPGA의 HPS (hard processor system)로 내장되어 있다는 점을 고려한다면 하드웨어 가속기가 ASIC으로 구현된다면 더욱 높은 성능 향상을 얻을 수 있을 것으로 예상된다.

압축된 음성 데이터를 인터넷 네트워크로 전송할 때 걸리는 시간을 비교한 결과는 다음과 같다.

구성	압축 + 전송 시간
기본구성 (압축 없음)	692 ms
소프트웨어 압축	568 ms
하드웨어 압축 (가속기)	552 ms

제안한 음성 압축 가속기를 적용할 경우 기본 구성 대비 25%의 네트워크 지연시간을 줄일 수 있다. 소프트웨어 압축 방식보다는 약 4%정도의 지연시간 이득이 관찰된다.

V. 결론

본 논문에서는 인공지능 스피커의 반응 시간을 개선하기 위한 방법으로 FPGA SoC에 하드웨어 음성 압축 가속기를 구현하여 적용하였다. 간단한 음성 압축 알고리즘을 FPGA에 구현하였으며, 이러한 방식은 소프트웨어 적으로 구현하는 것 대비 성능 우위를 가지고 있음을 실험으로 증명하였다. 네트워크 지연시간까지 고려할 경우 제안된 음성 압축 가속기를 통하여 약 25%의 네트워크 지연시간을 줄일 수 있다. 만약 열악한 무선 네트워크 환경이나 트래픽이 심한 환경이라면 음성 압축 가속기를 통하여서 인공지능 스피커의 네트워크 지연시간을 더욱 효율적으로 감소시킬 수 있을 것이다.

참고문헌

[1] Paul Garten, Alexa User Guide, Amazon Digital Services LLC - Kdp Print Us, 2019

[2] International Telecommunication Union, Pulse code modulation (PCM) of Voice Frequencies,ITU-T G711, 1988

[3] Texas instruments, A-Law and mu-Law Companding Implementations Using the TMS320C54x, 1997

[4] www.terasic.com

[5] www.intel.com

[6] Altera, External Memory Interface Handbook, 2016

[7] Intel Corporations, DE1-SoC Computer System with ARM Cortex-A9, 2016

[8] Intel Corporations, Using Linux on DE-series Boards, 2019

[9] Cornell University, DE1-SOC Docs ECE 5760, 2017

[10] www.aws.amazon.com

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구입니다. (No. 2019-0-00533, 컴퓨터 프로세서의 구조적 보안 취약점 검증 및 공격 탐지대응)