실시간 딥러닝 음성 신호 처리를 위한 하드웨어 구현

이성룡, 신예린, *유호영
충남대학교 전자공학과
e-mail : *srlee.cas@gmail.com, yrshin.cas@gmail.com, hyyoo@cnu.ac.kr*

Hardware Implementation for Real-time Deep Learning Speech Signal Processing

Sungryoung Lee, Yerin Shin, *Hoyoung Yoo
Department of Electronics Engineering
Chungnam National University

Abstract

Recently, deep learning is widely used to provide more accurate and superior performance compared to the traditional processing. In this paper, we describe a prototype configuration for real-time deep learning especially for speech signals. Since speech signal demands time to be digitalized, the proposed prototype configuration is carefully designed to satisfy the real-time operation.

I. 서론

최근 인공지능 기술이 발전함에 따라 다양한 분야에 적용되고 있다. 이미지, 얼굴인식, 자율주행뿐만 아니라 디바이스와 인간 간의 인터페이스 기술 중 인간의 커뮤니티케이션 형태와 가장 유사한 음성인식 기술에도 딥러닝 기술이 적용되고 있다. 하지만 음성 인식 딥러닝 처리과정은 복잡하고 시간이 오래 걸려 고성능이 아닌 디바이스에는 탑재에 어려움이 존재한다. 이런 이유로 서버형 기반의 임베디드 플랫폼 방식의 음성인식 기술로 발전하고 있으며 한정된 자원에서 다양한 서비스로 지원하는 형태로 개발되고 있다[1].

본 논문에서는 실시간 딥러닝 음성 신호 처리를 위한 하드웨어를 구현한다. 실시간 딥러닝 음성 신호 처리를 위해 사용된 하드웨어와 소프트웨어의 구성을 설명한다. 실제 구현을 통해 실시간 음성 처리에 대한 제약조건을 살펴보고 오실로스코프를 통해 동작 검증을 진행한다.

II. 본론

2.1 하드웨어 구성

실시간 딥러닝 음성 신호 처리를 위한 하드웨어는 음성 신호를 송수신하는 코덱과 딥러닝 기반의 데이터 처리하는 프로세서로 구성된다.

코덱은 그림 1과 같이 음성을 녹음하는 마이크, 아날로그/디지털로 음성 처리하는 ADC/DAC, 프로세서와의 통신을 위한 오디오 인터페이스로 구성된다. 음성 신호가 들어오게 되면 앰프에 의해 소리를 증폭하게 되고 사용자가 설정한 샘플레이트(Sample rate)와 비트수 그리고 채널(모노/스테레오)에 맞춰 데이터를 음성신호로 변환하게 된다. 변환된 데이터는 오디오 인터페이스를 통해 프로세서로 데이터를 전달하게 된

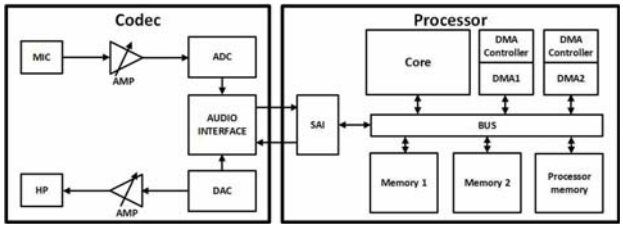

그림 1. 실시간 음성 신호 처리용 하드웨어 구성

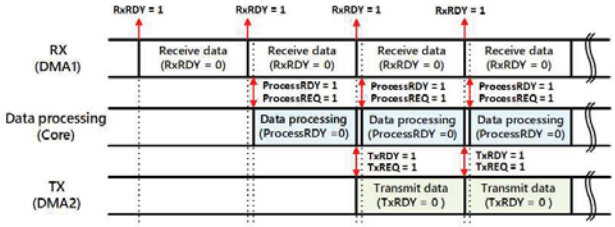

그림 2. 실시간 음성 신호 처리용 소프트웨어 구성

다. 오디오 인터페이스를 통해 수신된 데이터는 DAC를 거쳐 아날로그 신호로 변환하게 되고, 음성으로 출력하게 된다.

프로세서는 그림 1과 같이 Core 기반의 각 주변기기들이 붙어있는 버스 플랫폼 형태로 구성된다. 사용된 주변기기로는 메모리, DMA(Direct Memory Access), SAI(Serial Audio Interface)로 구성된다. DMA는 제어 신호를 인가 시 내부의 컨트롤러에 의해 Core 동작과 독립적으로 메모리에서 주변기기로(또는 그 반대로) 처리할 수 있어 실시간 데이터를 처리할 때 사용된다[2]. 만약 프로세서가 코덱 통합형 칩이 아닌 경우에는 그림 1와 같이 코덱과 분리된 칩으로 구성해야 되며, SAI 프로토콜을 이용하여 인터페이스를 구축해야 된다[3]. SAI 프로토콜은 I2S, PCM/DSP, TDM, AC'97, LSB/MSB-justified 프로토콜을 지원하여 사용자의 선택에 따라 원하는 프로토콜로 변경 가능하다[3]. 본 논문에서는 그림 1과 같이 두 가지 하드웨어로 딥러닝 음성 신호 처리할 수 있도록 구성하였다.

2.2 소프트웨어 구성

동작 소프트웨어는 그림 2와 같이 총 3단계의 동작으로 구성된다. 음성 데이터 수신단계(RX), 딥러닝 기반의 음성 데이터 처리단계(Data Processing), 처리된 음성 데이터를 코덱으로 송신하는 단계(TX)로 구성된다. 실시간으로 음성 처리하기 위해서는 2가지 제약조건이 존재한다. 1. 각 동작 단계는 Core와 독립적으로 연속 동작해야한다. Core가 모든 동작 단계를 처리하게 되면 처리하는 단계마다 다른 단계에서는 멈추는

그림 3. 오실로스코프로 동작 검증한 파형

문제가 발생한다. 이런 이유로 각 단계를 Core와 개별적으로 처리하는 컨트롤러가 필요하다. 2. 식 1에 기반한 DMA 처리시간 내 처리하는 딥러닝을 사용해야 한다. 음성 데이터를 처리할 때 Core와 DMA는 독립적으로 진행된다[2]. 각 처리시간의 타이밍이 맞지 않을 경우 데이터 충돌이 일어날 수 있다. 예를 들어 샘플레이트 16kHz, 데이터 전송 프레임(frame)이 128 일 경우, 데이터 처리 시간은 1/16000 * 128 = 4ms 시간 내에 처리가 되어야 하며 그 이상의 시간으로 동작할 시 데이터 충돌이 일어날 수 있다.

$$DMA\ processing\ time = \frac{1}{(Sample\ rate)} \times frame \quad ,(1)$$

Ⅲ. 구현 및 결론

실시간 딥러닝 음성 처리 시스템 구현은 대중적으로 사용되는 샘플레이트 16kHz, 비트수 32bit, 채널을 스테레오(2ch)로 설정하였으며, 내부에 전송 프레임 128, 두 개의 DMA 그리고 시간 내 처리하는 딥러닝을 사용하여 설계하였다. 오실로스코프로 검증한 결과, 그림 3과 같이 실시간으로 음성 처리할 수 있음을 확인하였다. 이를 통해 서버가 아닌 임베디드 기기로 처리할 경우 딥러닝 처리의 경량화가 필요하며, 경량화를 할 수 없는 경우에는 하드웨어 솔루션이 필요하다.

참고문헌

[1] 한국전자통신연구원(ETRI), 딥러닝 기반의 서버형 음성인식 기술, 2018.

[2] STMicroelectronics, AN2548: Using the STM32 F0/F1/F3/G0/Lx Series DMA controller, 2020.

[3] STMicroelectronics, STM32L4-Peripheral-Serial-Audio-Interface(SAI), Revision 3.1.