임베디드 환경에서의 인공지능 기반 음성인식 시스템 전력 소모 패턴 분석

박상기, 박상수, 정기석*
한양대학교

wdcup20002@hanyang.ac.kr, sonicstage12@naver.com, *kchung@hanyang.ac.kr

Analysis of Power Consumption for AI based speech recognition on Embed system

Park, Sang Ki, Park, Sang Soo, and Chung, Ki Seok*
Hanyang University, Seoul, Korea

요 약

최근 인공신경망 기반의 음성인식 시스템을 적용한 솔루션들이 많이 나오고 있다. 대부분의 솔루션들은 데이터 처리를 서버상에서 하고 사용자 단말에서는 결과값을 바탕으로 동작을 하는 형태이다. 네트워크가 연결되지 않은 독립적인 임베디드 환경에서 이를 구현하기 위해서는 제한된 전력의 소모 패턴 분석에 기반한 효율인 전력 관리가 필요하다. 본 논문에서는 임베디드 보드 환경에서 음성인식 솔루션을 실행했을 때의 전력 소모 패턴을 분석하였다.

Ⅰ. 서론

최근 스마트폰의 음성 기반 개인비서 서비스와 스마트 스피커 및 동시 통역 등 음성 인식 기반의 인공지능 서비스가 늘고 있다. 이런 음성 데이터의 처리를 위해서는 음향모델과 언어모델의 구축이 필요하다.

음향모델의 구축으로는 Hidden Markov Model (HMM)과 DNN 을 사용하는 DNN-HMM 모델 등이 사용된다. 또한 인식된 음성의 의미를 파악하는 언어모델에서는 weighted Finite State Transducer (wFST) 방식이 기존 모델들에 비해 우수한 성능을 보여주고 있다.

현재 대부분의 인공지능 기반 음성인식 서비스는 네트워크를 통해 데이터를 서버로 전송하여 처리하고 사용자 단말에서는 처리 결과를 바탕으로 알맞은 동작을 하도록 되어있다. 임베디드 환경 내에서 모든 연산을 구현하려면 제한된 전력량이 문제가 된다. 대부분의 음성기반 서비스들이 주요 계산을 서버에서 하는 방식을 택하는 것이 그러한 이유이다. 그러므로 뉴럴 네트워크 실행에 있어서 전력소모의 패턴을 분석하여, 전력소모가 많은 부분을 중점적으로 최적화 할 필요가 있다. 이를 위해서, 본 논문에서는 DNN-HMM 과 음성 인식의 각 단계들을 알아보고, 각 단계별 전력 소모량을 분석하였다.

Ⅱ. 본론

2.1 음성인식의 구조

뉴럴 네트워크의 기본 단위는 노드 이다. 일반적인 노드는 n 개의 입력 값에 각각 서로 다른 가중치를 곱하여 더한 뒤, 활성함수를 이용해 최종 출력 값 y 를 결정한다. 음성인식은 뉴럴 네트워크를 통한 음향모델의 구축과, 분석된 음향의 의미를 이끌어 내는 언어모델의 두가지로 이루어 진다. 데이터가 많이 모일수록 좋은 성능을 보인다. 서비스의 형태에 따라 임베디드형과 서버형으로 나누며, 키워드 기반, Large Vocabulary Continuous Speech Recognition (LVCSR) 기반의 결과 출력 형태가 있다[1].

2.1.1 DNN-HMM 의 구조

DNN-HMM 은 기존의 Gaussian Mixture Model (GMM)을 이용한 GMM-HMM 을 대체하는 모델이다. 연산량이 증가하지만 보다 정교한 확률 모델링이 가능하다. 그림 1 은 DNN-HMM 의 구조도를 보여준다[2].

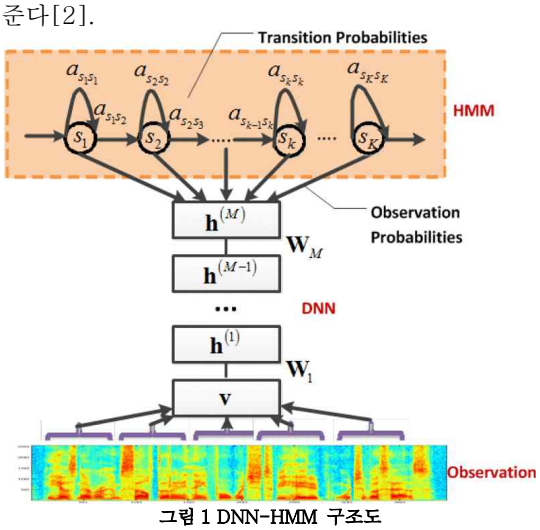

그림 1 DNN-HMM 구조도

2.1.2 언어모델의 구조

언어모델은 단어 순서에 대한 확률 분포도 이다. 단어 간의 전이를 확률로 표현하여 최종적으로 의미를 파악한다. Finite State Network (FSN) 또는 N-gram 모델이 많이 사용된다. 네트워크 변경이 자유로운 동적 탐색 형에는 Flat lexicon, Lexical tree 등이 있으며, 정적 탐색형에는 weighted Finite State Transducer (wFST)가 있다[3].

2.2 전력 소모 측정 환경

전력 측정은 NVIDIA Jetson TX2 임베디드 보드에서 진행하였다. Jetson TX2 는 GPU 로는 Pascal 기반의 256 CUDA 코어가 있으며, CPU 로는 Denver 코어 2 개와 ARM A57 코어 4 개로 구성되어 있으며, 모두 활성화 한 상태에서 전력 소모를 측정하였다. 전력 측정은 보드에 기본 내장된 PMIC 를 이용하여 CPU, GPU, 그리고 DDR4 메모리의 전력 소모량을 각각 측정하였다. 실행 음성인식 프로그램은 KALDI [4] 프레임워크를 이용한 것으로, 특징 추출 (feature extraction)은 가장 널리 사용되는 방법인 MFCC (Mel-Frequency Cepstral Coefficients)를 사용하였다 [5] 뉴럴 네트워크는 429 input dimensions 과 6063 output nodes 로 이루어진 4-layer DNN 구조이며, 언어 모델은 약 8 만개의 단어 데이터가 있는 wFST 를 사용한다. 실험용 프로그램은 미리 학습된 네트워크과 언어 모델을 사용하여 학습(training) 과정은 생략되었으며 추론 (inference)만 실행하면서 전력 소모를 측정하였다. 테스트는 총 길이가 43 분정도의 400 개의 음성데이터를 사용하였다.

2.3 전력 소모 측정 결과 및 분석

그림 2 는 MFCC 에 기반한 특징 추출 단계에서의 전력 소비량이다. 앞부분 약 20 초는 프로그램 실행 전 단계, idle 상태이므로 그래프에서 제외하였다. Idle 상태에서는 총 3W 였던 전력 소비량이 최대 11.2W 까지 올라가게 된다. 각 하드웨어 구성요소 별 시간차이는 있지만 전력 소모의 패턴은 비슷한 것을 확인할 수 있다. 40 초간의 실행에서의 평균 소비전력은 7.82W 이다.

그림 3 은 DNN-HMM 뉴럴 네트워크의 추론 및 wFST 모델 연산에 대한 전력 소비량이다. MFCC 과정과는 달리, CPU 소모전력이 줄어드는 부분 바로 뒤에 GPU 와 메인 메모리의 소모 전력이 일정 시간 상승하는 것을 볼 수 있다. 특히 GPU 는 전력 소모의 변화량이 큰 편이다. 최대소모 전력은 5.371W 였고, 최소 소모 전력은 0.228W 였으며, 그 차이는 5.142W 에 달한다. 100 초간 평균 소비전력은 8.73W 이다. 실험에 사용된 프로그램을 수행하는데 있어 총 140 초 동안의 평균 전력 소모는 8.47W 이다.

위의 두 전력 소모의 결과를 통해서 뉴럴 네트워크 기반의 음성인식 프로그램에서 연산량이 증가될 때의 전력 소모 패턴은 CPU 가 연산량이 과도한 작업을 GPU 에게 할당해주고, 그에 따라 다량의 메모리 접근을 통한 데이터의 이동이 발생하는 것이 반복적으로 나타나고 있음을 알 수 있다. 구현된 뉴럴 네트워크가 4-layer DNN 과 1 output layer 로 구성된 것이고, 4W 가 넘는 GPU 사용량이 보이는 것이 5 개 인 것으로 보아 각 층의 연산마다 전력 소모량이 증가하는 것으로 보인다. 이러한 전력 소모 패턴 분석은 추후 CPU 와 GPU 의 통합적 전력 관리를 통한 임베디드 플랫폼에서의 음성 인식 솔루션의 전력 소모 최적화에 유용할 것이 기대 된다.

Ⅲ. 결론

임베디드 환경에서 인공신경망 기반의 음성인식 시스템 개발에 대한 관심이 높다. 임베디드 환경에서 많은 연산량을 처리하게 되면 전력 소모 문제가 가장 심각해진다. 전력 소모를 최소화하기 위해서는 정확한 전력의 소모 패턴의 모니터링 및 분석이 필요하다. 본 논문에서는 임베디드 보드 환경에서 음성인식 솔루션을 실행했을 때의 전력 소모 패턴을 분석하였다. 인공지능 신경망의 구성에 따른 CPU 와 GPU 에서의 전력 소모 패턴 분석을 통해 효과적인 전력 관리가 가능할 수 있음을 보였다.

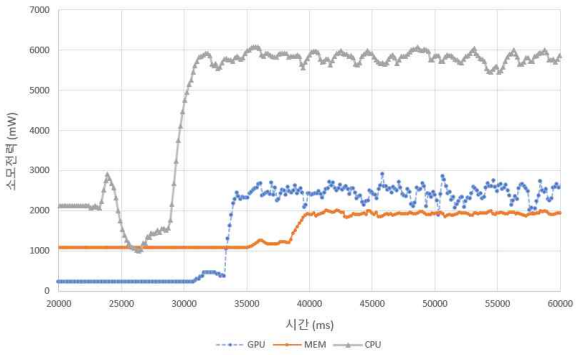

그림 2 특징 추출 단계의 전력 소모

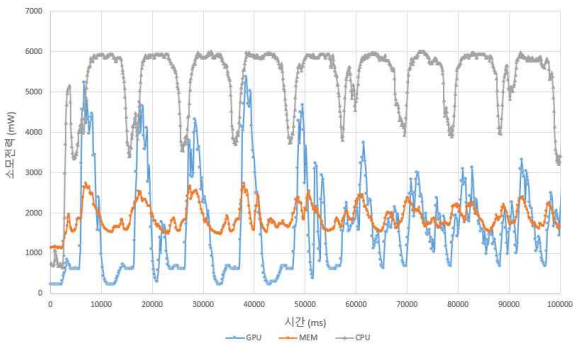

그림 3 디코딩 단계의 전력 소모

ACKNOWLEDGMENT

이 논문은 2015 년도 정부 (교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업 임 (NRF-2015R1D1A1A09061079)

참 고 문 헌

[1] George E. Dahl, Dong Yu, Li Deng, Alex Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMS," ICASSP, Prague, pp.4688-4691, 2011

[2] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30-42, Jan. 2012.

[3] M. Mohri, F. Pereira, M. Riley, " Weighted finite-state transducers in speech recognition " Computer Speech and Language, pp. 69-88, Jan. 2002

[4] P. Daniel, G. Arnab, B. Gilles, et al., " The Kaldi speech recognition toolkit", IEEE Signal Processing Society, Hawaii, 2011 (http://kaldi-asr.org/)

[5] L. Muda, M. Begam, I. Elamvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", arXiv:1003.4083, Journal of Computing, Volume 2, Issue 3, March 2010