ETRI

2014
Electronics and
Telecommunications
Trends

소프트웨어 기술동향 특집

내장형 음성인식기를 위한 전용 하드웨어가속기 기술개발 동향

Trends of Hardware Accelerator for the Embedded Speech Recognition

김주엽 (J.Y. Kim) 통방융합 SoC 연구실 선임연구원
김태중 (T.J. Kim) SW-SoC 개방형플랫폼팀 팀장
이주현 (J.H. Lee) 통방융합 SoC 연구실 실장
엄낙웅 (N.W. Eum) 시스템반도체연구부 부장

사람의 말소리를 문자로 변환하여 기기의 제어명령으로 활용하는 것이 음성인식 기술이다. 음성인식에 대한 기술개발 요구는 수십 년 전부터 있어 왔고, 꾸준히 제품화되고 있는 분야라 하겠다. 제품으로의 상용화가 가능한 알고리즘 및 데이터 처리체계는 HMM(Hidden Markov Model)이라는 수학적 모델링으로 정형화되어 있으며, 대규모의 반복적 데이터 수집과 정교한 학습 데이터베이스의 구축이 음성인식기술의 핵심요소라는 것이 일반적인 시각이다. 이러한 이유로 인해, 대용량 음성인식 데이터베이스의 수집, 가공 등이 가능한 인프라를 갖춘 기관 및 업체들이 음성인식기술 시장을 점유할 수 있는 것이다. 그러나, 이러한 음성인식의 서비스 제공 체계는 사물인터넷 또는 웨어러블 디바이스 등으로 음성인식 사용자 인터페이스가 확대되고 통신 및 네트워크가 연결이 불가한 경우 그 한계를 보일 수 있다. 본고에서는 이러한 문제를 해결하기 위한 내장형 음성인식기의 하드웨어가속기 기술개발에 대한 내용과 국내외 현황을 살펴보기로 한다.

Ⅰ. 서론

본고에서 언급하고 있는 음성인식기술은 사람의 말소리를 텍스트로 변환하는 기술을 말한다. 사람의 말소리를 글자로 변환하는 기술은 키보드, 마우스와 같이 사람의 손동작에 의존한 방식과 달리 편의성 측면에서 많은 분야에 적용하기 좋은 기술이다. 특히, 최근에 출시되고 있는 착용형 기기들은 입력 버튼과 입출력 기기 연결의 제약으로 음성인식기를 탑재하는 것이 기본으로 인식되고 있다.

이러한 음성인식기술은 실험실 수준의 일정한 조건하에서 그 정확도가 거의 95% 이상 도달하고 있으며, 전투기와 같은 과도한 소음환경에서의 음성인식, 화자의 인식을 통한 음성인식 등 음성인식기에 대한 기술은 계속 진화하고 있다.

현재 적용되고 있는 음성인식의 내부 알고리즘 및 처리체계는 HMM(Hidden Markov Model)이라는 패턴인식 모델링을 채택하고 있으며, 상용화되고 서비스 되고 있는 대부분의 음성인식기는 이 모델링 체계 안에서 약간의 변형과 부가적인 처리방식을 도입하여 발표한 것이라 보면 되겠다. 최근에 머신 러닝 연구분야의 인기로 다양한 패턴인식 알고리즘에 처리과정을 도입하여 연구되고 있으나 상용화 수준에서는 HMM 기반의 음성인식기가 주를 이룬다.

실제로 개발자와 연구자들을 위한 참고 소스코드가 영국 캠브리지 대학의 HTK(Hidden Markov Model Tool Kit)와 미국 카네기 멜론 대학의 Sphinx 등이 공개되어, 개인용 PC에 다운로드 받아 개발자 매뉴얼 도움을 받으면 자신만의 초보적인 음성인식기 개발이 가능하다.

현재 대부분 성공적으로 서비스되고 있는 음성인식기는 사용자 단말기 내부에서 처리되는 것이 아니다. 원격지의 대용량 음성인식 데이터베이스를 가지고 있는 고성능의 컴퓨팅 기기에서 처리되어 그 결과를 받아 보는 형태이다. 이러한 처리체계는 두 가지 문제를 안고 있다. 첫째, 단말기와 원격지 고성능 기기 사이의 연결이 단절 된 경우 서비스가 불가하다. 둘째, 음성인식 서비스 요구가 증가 할 경우, 응답 시간의 지연 등의 문제로 사용자 편의성이 줄어 드는 문제가 발생할 수 있다.

이를 해결하기 위해서는 단말기 내부에 음성인식을 처리할 수 있는 엔진을 각자의 기기가 가지고 있는 경우 해결되나, 앞서 설명한 음성인식 학습 데이터베이스의 문제를 안고 있을 수 있어, 음성인식 학습 데이터베이스를 필요 시에 업데이트 할 수 있는 기능이 필요하다.

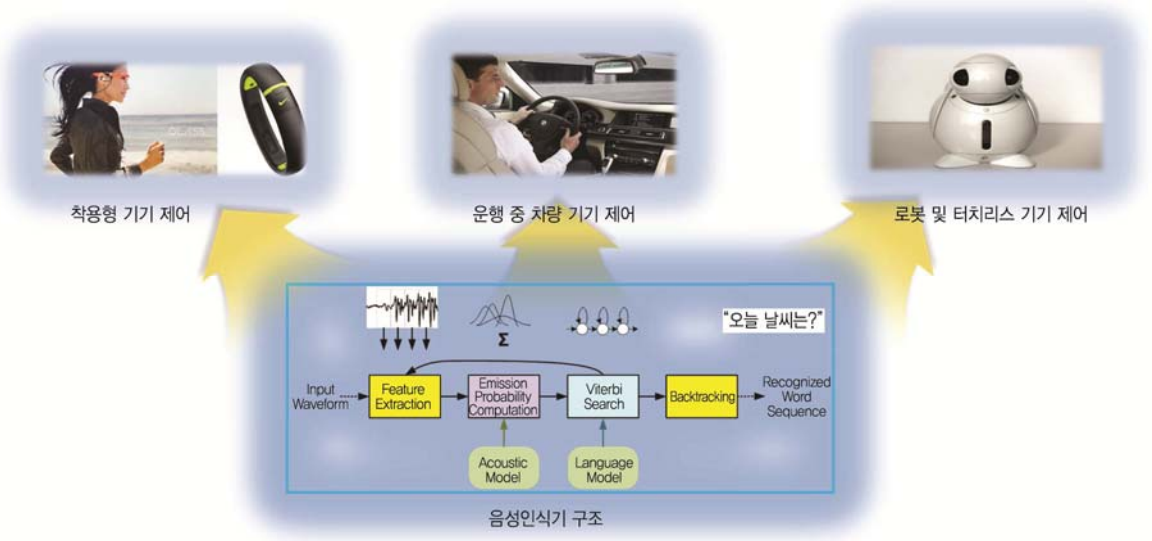
본고에서는 음성인식기를 기기 자체적으로 내장하는 경우 중에서 음성인식 하드웨어가속기 기술개발에 대한 내용을 다루기로 한다. 음성인식이 사용자 인터페이스로 일반화될 경우, 칩 내부의 전용 IP 또는 별도의 칩으로 구현하는 것이 필요하다.

이미, 인텔은 프로세서 내부에 음성으로 기기를 “Wake-Up”시킬 수 있는 기능의 음성인식 하드웨어 모듈을 삽입하여 발표하였으며, 세계적인 음성인식기 개발회사인 뉘앙스도 칩 또는 SoC(System On a Chip) 개발회사와 공동 개발 의지를 발표한 바 있다.

본고에서는 내장형 음성인식기의 필요성, 내부 구조 및 기능, 마지막으로 국내외 개발 및 적용사례를 소개하기로 한다.

Ⅱ. 내장형 음성인식기의 필요성

(그림 1)에 도시되어 있는 바와 같이 내장형 음성인식기의 활용분야는 최근에 그 필요성이 확장되고 있다. 착용형 기기의 대중화, 사물인터넷을 통한 소형기기의 제어와 사용자 편의환경 제공 측면에서, 음성인식 사용자 인터페이스는 일반화되고 있는 추세라 하겠다. 소형기기에 다양한 명령어를 구현하기 위해서는 오디오 입출력 기기를 활용하는 것 이외에는 필요하지 않아, 그 형



(그림 1) 내장형 음성인식기의 응용분야

상적 구현이 비교적 용이하다.

뿐만 아니라, 선진국 일부 국가에서는 자동차 운전 중에, 휴대폰 및 내비게이션을 조작하는 행위는 법으로 금지하여, 강력하게 처벌하고 있는 추세라 운전 중에 음성을 통한 기기 제어는 자동차 전장 시스템에 필수적인 기술로 자리매김하고 있다. 이러한 이유로 인해, 최근에 국내외 출시되고 있는 대부분의 자동차 내비게이션 및 오디오 기기에는 음성인식 인터페이스 기능이 내장 되어 있다.

요구하는 비행기 조종석에서의 레버 컨트롤을 음성을 통한 직관적 설정이 가능하도록 최신 전투기와 민간 항공기에 적용되어 이미 상용화 되어 있으며, 손 동작을 통한 명령어 입력이 어려운, 주방, 수술실, 운동 기기에 대한 제어 및 명령어 입력을 위한 기술이 음성인식기를 적용한 사용자 인터페이스 기술이라 하겠다.

이러한 음성인식기에 대한 구현 방식은 소프트웨어적인 엔진을 개발하는 것에서부터, 하드웨어적인 가속기를 만드는 것까지 다양하다. 적용 시스템의 형상 및 사양 등에 따라 그 구현 형태의 유리함과 불리함을 따질 수 있으나, 일반적으로 정형화된 음성 패턴, 고립어 인식 같은 경우에 하드웨어 구현이 더 효과적이라 하겠다. 특히, 기기의 저전력화 및 주 프로세서의 오프로드 측면에서 음성인식 처리를 위한 전용 하드웨어를 코프로세서나 IP화 하는 것이 효율적이라 하겠다.

특히, 기가가 저전력 모드에서 비 접촉식 동작으로 "Wake-Up" 기능을 실현하기 위해서는 항상 마이크로 입력되는 소리에 대해서 음성인식을 수행해야 하는 문제를 안고 있다. 이 때문에, 저전력화를 실현할 수 있는 전용 하드웨어 음성인식기로의 구현이 효과적이라 하겠다. 또한 복수의 어레이 마이크 입력을 통한 소음 제거 및 반향 제거 등 비교적 계산 복잡도가 증가하는 경우에도 전용 하드웨어를 활용하는 것이 필요하다.

Ⅲ. 음성인식기의 신호처리

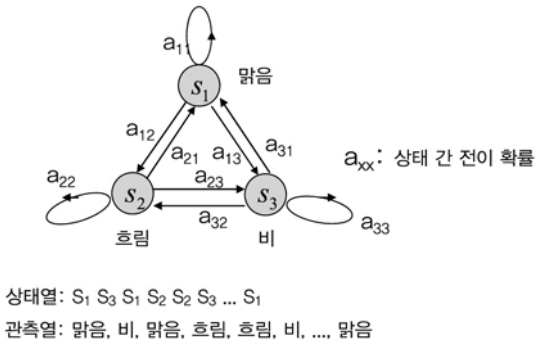
1. HMM 기반의 확률 모델

확률 및 통계적 수학적 모델 중에서 사전 확률으로 사후 확률을 예측하는 방법을 기술한 것이 베이즈 이론이다. 베이즈 이론을 통해 우리는 사전에 발생한 사건의

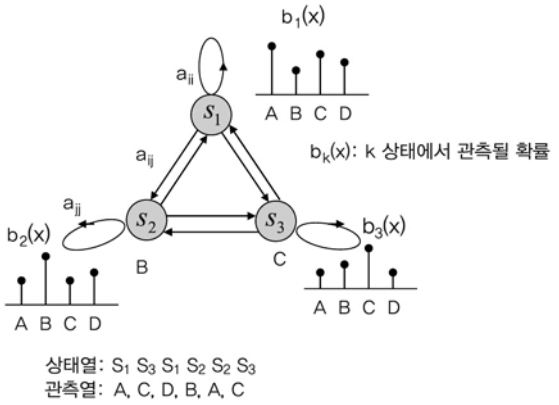
조합과 그 확률 값을 통해 앞으로 발생할 사건들에 대한 가능성을 예측 할 수 있다. 이러한 베이즈 이론을 기반으로 여러 사건들의 발생 확률 관계를 정리하는 것이 마르코프 모델(Markov Model)이다. 마르코프 모델을 통해 사건의 상태를 뜻하는 원과 사건과 사건 사이의 변화를 전이 화살표로 표현하면, 예측하려는 확률 모델을 유한상태머신(Finite State Machine)과 같이 시각화하여 표현할 수 있다.

날씨상태를 마르코프 모델로 나타내면 (그림 2)와 같이 시각화 할 수 있다. 그래서 초기 날씨의 확률값, 그리고 지금까지 관측된 날씨상태의 변화를 기반으로 생성한 조건부 확률값을 가지고 앞으로의 날씨변화를 예측하는 확률 모델을 만들 수 있는 것이다.

이와 같이, 관측된 값이 바로 특정 사건이나 상태를 뜻하는 것이라면, 마르코프 모델을 이용하여 확률값을 추출하는 것이 가능하다. 그러나 관측된 값이 특정 사건이나 상태를 대표하는 것이 아니라, 특정 사건이나 상태에 대한 가능성만을 알려 준다면, 일련의 관측치 나열이 어떤 상태의 나열과 동일한지에 대한 문제에 직면하게 된다. 이러한 문제를 해결하기 위해 고안된 것이 HMM (Hidden Markov Model)이다. 마르코프 모델은 사전에 발생한 사건과 상태들의 나열을 통해 다음 사건 또는 다음 상태를 예측하기 위한 수학적, 확률적 모델이었다면, HMM은 일련의 관측값을 통해 상태 또는 사건의 나열을 찾아가는 작업을 수행하는 문제로 옮겨가게 된다((그림 3) 참조). 이러한 문제에 대한 수학적 내용은 1960년대 초반부터 이루어졌고, 수학 및 확률문제를 연구하는 학자들에 의해서 이론적 토대가 구축되었다.



(그림 2) 날씨에 관한 마르코프 모델 예

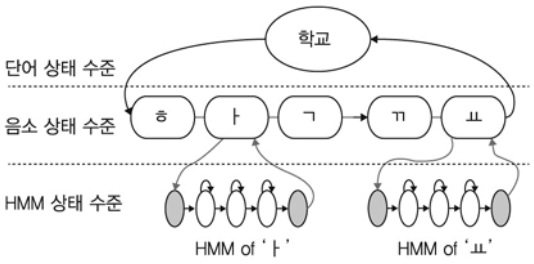


(그림 3) HMM 모델의 도식 예

HMM이 음성인식분야에 본격적으로 적용된 것은 Rabiner 교수가 발표한 논문부터라고 말할 수 있다[1]. 그의 논문은 HMM 기반으로 발표된 음성인식기의 대표적 인용 논문이며, 대부분의 HMM 설명자료에서 언급되며, 음성인식분야를 연구하는 대부분의 사람들이 완벽히 이해해야 한다.

사람의 말소리를 듣고 글로 변환하는 작업은 사전에 말소리를 글로 변환할 수 있는 사전지식을 가지고 있기 때문에 가능하다. 사전지식이라고 하면, “학교”라는 말소리에 대한 음향학적 또는 언어학적 사전학습이 토대가 되어 가능하다는 뜻이다. “학교”는 음향학적으로 구분 가능한 소리 수준인 “ㅎ”, “ㅏ”, “ㄱ”, “ㅛ”, “ㅛ” 로 분리 가능하며, 우리는 이를 언어학적으로 음소라고 칭하고 음소의 나열을 글자로 변환시킬 수 있는 학습을 반복적으로 수행하여, 말소리를 문자로 이해하고 적는 것이 가능하다.

이러한 사람의 말소리에 대한 이해 능력은 HMM을 빗대어 설명이 가능하다. 사람들은 자신들이 이미 학습한 소리와 문자 관계를 음향학적 또는 언어학적 연결관계를 머릿속에 가지고 있다. 그래서, “학교”에 대한 말



(그림 4) 음성인식의 계층적 확률 모델

소리를 입력으로 "ㅎ", "ㅏ", "ㄱ", "ㄲ", "ㅛ" 순서의 소리를 듣게 되면, 자신이 가지고 있는 학습정보와 가장 가능성이 높은 단어를 문자로 떠올리게 된다. "ㅎ"을 듣고 "ㄱ"으로 인식하지 않고, "ㅏ"를 듣고 "ㅗ"로 이해 하지 않는 것은 자신이 들어 왔던 정보를 기반으로 가능성을 추출하는 확률 모델링을 머릿속에 구축하고 있기 때문에 가능하다. 더욱이 순차적으로 발생하는 음소의 관계를 통해서 예측의 정확도를 높이도록 되어 있고, 단어 단위로 구분하여 이러한 확률 모델링이 구축 되어 있는 것이다. 이를 단어, 음소, HMM 모델로 계층적으로 나타내면 (그림 4)와 같다.

위 도식에서의 HMM 상태들에서의 입력 패턴에 대한 관측 확률을 구하여 가장 누적 확률치가 높은 상태열이 음성인식 결과가 된다. HMM 상태열의 조합은 음소 상태열과 단어 상태 열과 연결 되는 계층화로 음성인식 학습 데이터베이스가 구축되는 것이다.

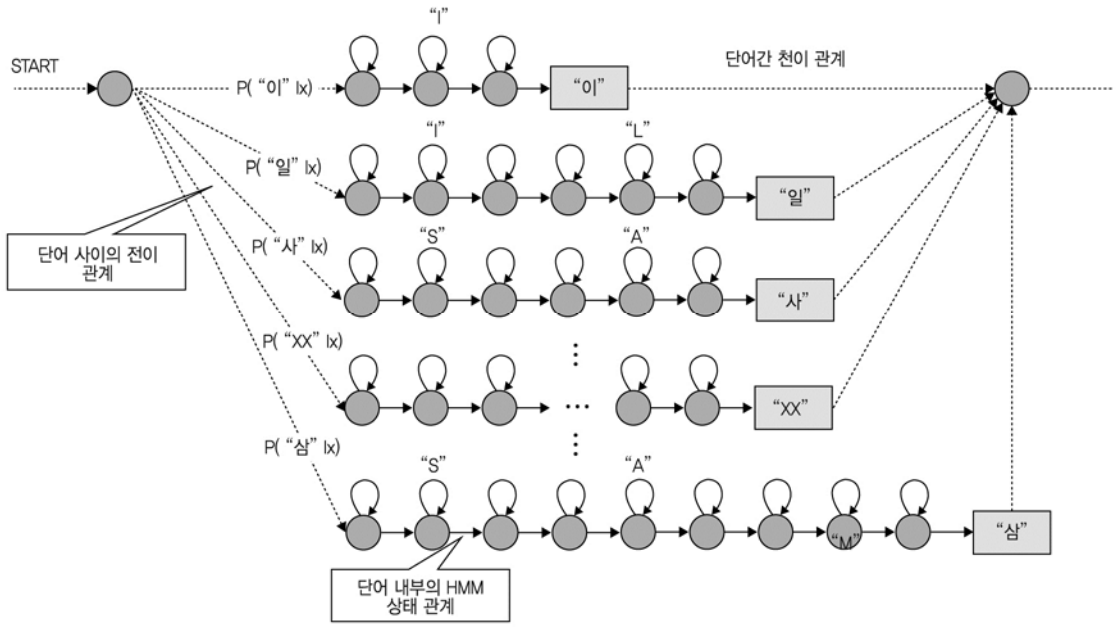
(그림 4)에서와 같이 단어 상태 수준과 음소 상태 수준에서의 상태라 칭하는 것은 그 구분이 이해 가능한 추상화 수준의 단위이다. 단어와 음소라는 단위로 상태들이 나뉘어져 있기 때문이다. 그러나 HMM 상태 수준에서의 상태는 음소 하나에 처음과 끝의 상태 수준을 제외하고 일반적으로 세가지 상태의 연결로 모델링 된다. 이때의 상태들은 사람이 특정 음소를 발성할 때, 필요한 패턴들의 값이 확률적으로 일관성을 유지하는 구간이라 생각하면 된다.

앞서 설명에서, HMM은 관측되는 표현값이 특정 상태와 바로 연결되는 것이 아니라고 하였다. 음성 인식에서 입력되는 음성패턴이 HMM 상태를 바로 표현하는 것이 아니라, 단지 가능성과 확률값을 얻어 낼 수 있다는 말이다. 이를 위해서 HMM 상태들은 입력 패턴값에 대한 확률값을 생성할 수 있는 확률밀도 함수에 대한 평균과 표준편차를 가지고 있어야 한다. 일반적인 확률밀도 함수인 가우시안 확률분포 함수를 기반으로 하고 있기 때문이다. 이것이 학습 데이터베이스를 구성하는 핵심 데이터가 되는 것이다.

HMM 상태들에 대한 확률밀도 함수를 기술할 수 있는 평균과 표준 편차들이 학습 데이터의 대부분을 차지하는 것이다.

그렇다면, 우리는 음소라는 단위로 문자화 할 수 있는 모든 말소리를 표현할 수 있으므로, 한글의 경우 40개 정도의 음소에 대한 HMM 상태들의 평균과 표준편차 파라미터만 가지고 있으면 음성인식이 가능할 것이라는 생각을 해 볼 수 있다. 그러나 실제 음성인식기의 HMM 상태들은 "학교"의 "학"에 종성음 "ㄱ"과 "학생"의 "학"에 종성음 "ㄱ"은 다른 음소로 취급한다. 실제적으로, 음성인식기는 인식을 위한 단어들을 구성하는 음소들은 각각 다른 음소로 취급하는 것이 원칙이다. 다시 말해, 독립된 단어 사이에 동일한 표식의 음소가 있더라도 다른 음소로 취급하여, HMM 상태들에 대한 확률밀도 함수의 파라미터들을 모두 저장하고 있어야 한다는 뜻이다. 이러한 원칙은 음소의 표식이 동일하더라도, 다양한 음소의 조합으로 발성되는 단어 내부에서는 제각각 다른 소리로 들리기 때문에, 정확한 음성인식을 위해서는 이러한 원칙을 준수해야 한다. 구현과 계산의 복잡성으로 인해 음성인식률을 고려한 최적화 기법 등이 적용되고 있으나, 기본적인 원칙은 이 방식에서 크게 다르지 않다.

그러므로, 음소를 기반으로한 음성인식이 아니라, 미리 설정된 단어들을 기반으로 음소가 나뉘어지고, 각각의 음소에 대한 HMM 상태들에 대한 입력 패턴의 확률값을 모두 생성하여 최적화 상태열의 조합을 찾는, 반복



(그림 5) HMM 상태 탐색 네트워크

작업을 수행하여 최고의 누적확률 값을 찾는 것이 음성인식기의 대략적인 계산과정이다.

이러한 HMM 알고리즘의 계산과정을 추상화하면, 단어 내부와 단어 사이를 HMM 상태들에 대한 탐색 네트워크(Search Network)로 도식할 수 있다((그림5) 참조). 즉, HMM 기반의 음성인식기는 인식하려는 단어들의 조합으로 도식화 할 수 있는 HMM 상태 사이의 최적화 경로를 찾는 알고리즘으로 정형화할 수 있음을 의미한다. 이러한 정형화 형태는 비터비(Viterbi) 탐색 알고리즘과 유사하다. 프레임마다 상태 관측 확률과 전이 확률의 조합으로 최적경로를 추출하는 방식은 비터비 알고리즘의 계산과정과 다르지 않다.

그러므로, HMM 기반 음성인식은 관측값에 해당하는 음성인식 패턴 생성, HMM 상태들에 대한 관측 확률의 생성, 탐색 네트워크 상의 최적경로를 찾는 비터비 탐색으로 요약할 수 있게 된다.

2. 음성인식 패턴 생성

HMM 알고리즘을 기반으로 하는 음성인식기는 말소리에 대한 ADC 데이터를 음성인식기의 입력 패턴으로 사용하지 않는다. 사람의 말소리를 통해 문자로 변환하기 위해서는 발성되는 음파에서 샘플링된 수준의 dynamic range의 데이터 크기는 필요하지 않다는 뜻으로 이해할 수 있다.

즉, 사람의 말소리에 대한 이해도는 음파의 샘플링된 데이터를 기반으로 결정되는 것이 아니라, 샘플링된 데이터를 재가공한 데이터가 필요하다는 뜻이다.

시간영역을 주파수 영역으로 변환하여 얻은 스펙트럼에 대한 피크값들이 말소리 이해를 위한 핵심성분이라는 것은 이미 음향학 분야에서 정의되어 있다.

다시 말해, 사람의 성도에서 공명현상으로 발생하는 배음들의 조합으로 만들어지는 스펙트럼의 피크값들을 칭하는 포먼트(Formant) 성분들의 분포양상으로 음소를 이해할 수 있게 되는 것이다.

포먼트와 유사한 패턴을 생성하는 방식이 캡스트럼이라는 신호처리 방식이 있다. 캡스트럼은 말소리 데이터를 푸리에 변환을 통해 얻은 주파수 영역값들에 로그값을 얻는다. 이 로그값들에 대한 역 푸리에 변환을 통해

얻은 것이 캡스트럼이다. 캡스트럼값은 사람이 인지할 수 있는 말소리의 신호적 특성을 그대로 모사하기에 가장 적합한 패턴이라는 것이 일반적인 분석이며, 말소리의 구분에 필요하지 않은 많은 신호 성분들을 제거하여 데이터량을 압축하는 효과도 얻을 수 있다.

오랫동안 음성인식 패턴 생성에 대한 연구가 진행되어 단순한 캡스트럼 생성에서 확장하여, 주파수 스케일을 사람 청각의 민감도를 반영한 MFCC(Mel-Frequency Cepstrum Coefficient)으로 확장할 수 있다. 이는 캡스트럼 과정 중에 있는 푸리에 변환 출력값에 사람의 청력 민감도와 유사한 필터 뱅크를 삽입하여, 주파수 스케일을 재조정하는 방법을 사용하는 것이다.

MFCC 생성은 입력 말소리 ADC 데이터를 10ms~15ms 주기적 간격으로 30ms~45ms의 크기의 프레임을 취하여 생성하게 된다. 사람의 말소리의 유의미한 통계적 변화는 10ms~15ms 간격 사이에서 변화될 수 있고, 30ms~45ms 범위에서 음성학적 통계치가 유효하다는 것이 일반적이다.

그러므로, 하나의 프레임에서 생성된 MFCC가 음성인식기의 패턴 데이터가 되는 것이다. 30ms 타이밍 윈도우에서 생성되는 16KHz 샘플링 주파수에서는 480개의 입력 ADC 샘플 데이터가 필요하지만 MFCC 결과 생성되는 데이터 수는 대략 12~39개 정도이다.

이처럼 음성인식에 필요한 데이터량은 실제 음성의 샘플링 데이터량에 비해 훨씬 줄어 드는 압축효과를 얻게 된다.

3. 관측 확률 생성작업

HMM 기반의 음성인식 처리과정에서 HMM 상태의 입력 패턴에 대한 확률값을 생성하기 위한 확률밀도 함수가 필요하다. 일반적으로 복수의 가우시안 확률밀도 함수의 가중치를 곱한 합으로 아래와 같이 기술 되는 GMM(Gaussian Mixture Model) 확률밀도 함수를 적용한다.

$$f(x) = \sum_{k} w_k \lambda(\mu_k, \sigma_k{}^2)$$

단수의 가우시안 확률밀도 함수에 비하여 GMM은 다중의 확률 피크를 갖는 자연적 패턴에 대한 확률분포를 기술하는데 적합하다.

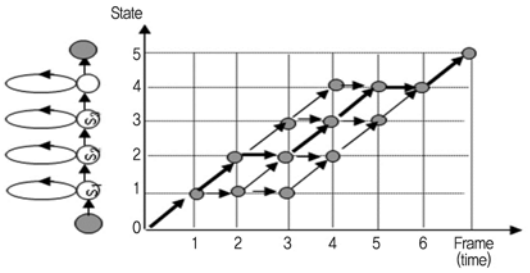
GMM을 통해서 얻은 확률밀도 함수는 앞서 설명한 HMM 상태들에서 관측값이 관측될 확률을 생성하는 역할을 한다.

HMM의 상태들은 음성인식을 위한 범위의 단어들에 속해 있는 음소별로 구분 되어 있으며, 음성인식 학습 데이터베이스 구축 시에 평균과 표준편차, 그리고 복수의 가우시안 함수에 대한 가중치가 미리 학습 데이터베이스 구축 시에 생성되어 메모리에 저장되어있어야 한다.

현재 입력 패턴값에 대한 모든 HMM 상태에 위치 하고 있을 가능성을 생성하여 저장하고, 탐색 네트워크 상에 최적경로를 찾는 비터비 탐색기에 넘겨 주어야 한다.

4. 최적경로 탐색과정

관측 확률 생성작업에서 얻어진 인식 단어들에 속한 모든 상태들에 대한 확률값을 저장 받고, 상태 간의 전이 확률은 이미 학습 데이터베이스로 가지고 있다. 그러므로, 앞서 설명한 것과 같이 비터비 탐색 알고리즘으로 최적경로를 찾으면 음성인식을 수행할 수 있다((그림 6) 참조). 10ms 간격으로 입력되는 음성 패턴에 대해 그림 5에 표시된 전체 탐색 네트워크의 최적경로를 찾아야 한다. 성능적으로 인식 단어 수가 증가할 경우 탐색해



(그림 6) 비터비 최적경로 탐색과정

야 하는 상태 수도 같이 증가하여 탐색시간이 증가하는 문제가 있어, 일정한 값 이하로 누적 확률값이 떨어지는 경로에 대해서는 인식결과로의 가능성이 없다는 것으로 간주하여, 다음 프레임의 입력 패턴에 대해서는 비터비 탐색후보에서 제외하는 기법을 활용하여, 탐색시간을 줄이기도 한다.

5. 성능 평가

음성인식기의 성능은 단어 오인식률(WER: Word Error Rate), 실시간 처리인수(RTF: Real Time Factor)라는 지수가 있다.

연속어 음성인식을 수행할 때, 음성인식기의 인식 결과와 본래의 음성 사이에 오차 즉, 단어의 추가, 상실, 치환의 오인식 결과를 모두 반영하여 지표로 나타낼 수 있다.

$$WER = \frac{S + D + I}{S + D + C}$$

S: 대체된 단어 수, D: 상실된 단어 수,
I: 삽입된 단어수, C: 인식된 단어 수

RTF는 음성을 인식하는 걸리는 음성인식 처리시간을 나타내는 것으로 입력된 음성의 입력시간 대비 처리시간을 뜻한다. RTF값이 적을수록 음성인식기의 실시간성이 증가하는 것으로 평가하면 된다.

Ⅳ. 국내외 개발사례

1. 국내 사례

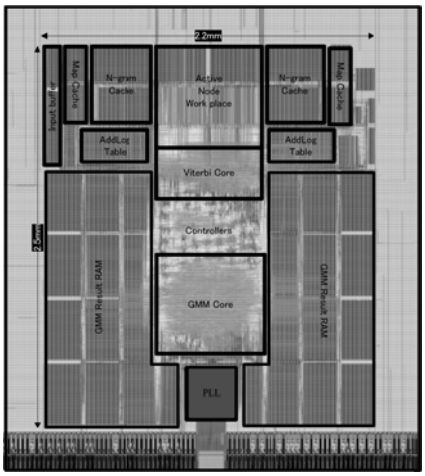
비교적 최근에, 국내에서는 서울대학교 연구팀에서 2만 단어 연속어 음성인식기를 하드웨어로 구현하였다. MFCC 생성과정을 제외한 관측 확률 연산과 비터비 탐색과정을 전용 하드웨어로 구현하여 FPGA(Field-Programmable Gate Array)에서 실험하여 학술지 등에 발표한 바 있다. 서울대학교 음성인식기의 특징으로써, 음성인식 학습 데이터베이스 접근에 대한 읽고 쓰는 작업에 대한 메모리 컨트롤러의 동작 최적화를 실현하였다. 이를 통해 100MHz에서 RTF 값을 약 0.4까지 동작하도록 구현하였으며, 음성인식률이 대략 92% 정도로 보이고 있다. 음성인식 학습 데이터베이스 확장으로 인식 단어 수를 높일 수 있도록 구조화된 것이 특징이다[2].

이외에 국내에서는, 과거 고립어 또는 단문의 음성인식기를 칩 또는 하드웨어로 구현하여 상용화한 경우는 일부 있었으나, 최근에 개발 및 연구사례로써 주목할만한 내용은 없다.

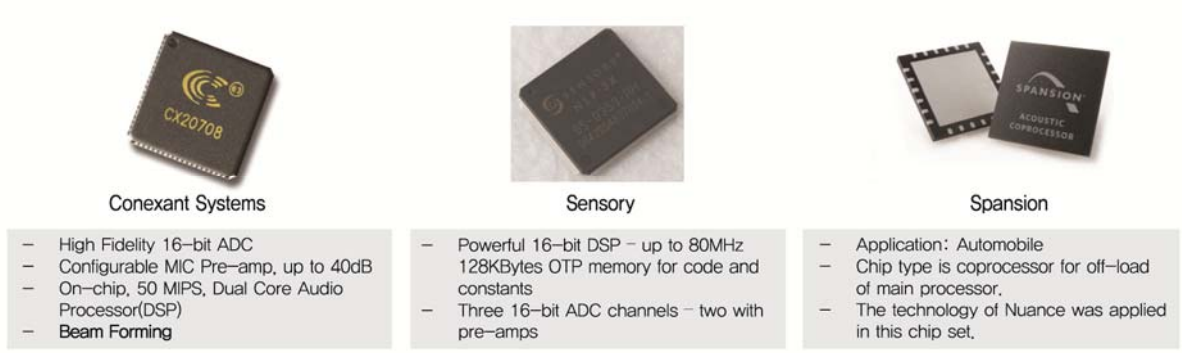
2. 해외 사례

일본 고베 대학 팀은 2012년도 학술지에 발표한 논문에서, 6만 단어 연속어 음성인식기를 40nm 공정에서 140mW 소모 전력을 갖는 칩을 구현하였다((그림 7) 참조). 저전력 동작을 실현하기 위해서 GMM 및 비터비 탐색기에 대한 최적의 파이프라인 동작구조를 실현하는 구조를 제안하고 있다[3]. 특히, 최근에 알려진 연구 및 개발사례 중, 음성인식 단어 수 대비 최소 전력소모를 기록하고 있다.

음성인식기를 칩으로 구현하거나 하드웨어로 개발하여



(그림 7) 고베대학 연구팀에서 개발한 칩[3]



(그림 8) 음성인식기 칩

판매하는 사례는 미국의 음향 및 음성신호 처리를 전문으로 하는 회사들에 의해서 주도적으로 이루어지고 있다.

미국의 Conexant Systems사는 복수의 음성 및 소리 신호 처리용 DSP 기반으로, 멀티 채널 오디오 ADC를 지원하여 빔포밍 및 고음질의 음성신호 처리가 가능한 칩을 판매하고 있다[4]. 또 다른 미국의 음성 및 소리신호 처리 칩 셋 개발 회사인 Sensory사도 음성인식 및 음성합성 칩 셋을 판매하고 있다[5]. 소음처리 및 화자 구분 및 인식 등의 기능적 특징을 내세우고 있다.

가장 최근에, 미국의 ROM 개발 회사인 Spansion사는 자동차 전장 부품에 활용 될 수 있는 음성인식 칩인 Acoustic Processor를 발표한 바가 있는데((그림 8) 참조), 음성인식 엔진 개발회사로 유명한 뉘앙스 소프트웨어를 탑재하고 있는 것으로 소개하고 있다[6].

앞서 설명한, 상용화된 음성인식 칩 셋 이외에 인텔은 음성신호에 대한 처리를 전용으로 하는 하드웨어 환경을 칩 셋 내부에 구현하여 프로세서의 로드를 줄이기 위한 노력과 동시에 항상 음성신호를 모니터링하며 음성을 인식할 수 있는 "Intel Smart Sound Technology"를 발표한 바도 있다[7].

V. 결론

음성인식을 위한 음성인식기는 학습 데이터베이스의 구축 및 개선을 위한 이유 때문에, 원격의 고성능 컴퓨팅 기기에 의존하는 방식으로 서비스되고 있는 실정이다.

이러한 방식이 음성인식률 측면에서 유리하나, 음성인식 서비스 요구의 증대와 통신장애 등으로 발생하는 문제 때문에, 내장형 음성인식기의 활용성은 여전히 유효하다. 특히, 웨어러블 기기와 자동차 기기에 대한 제어 등은 이미 내장형 자체 음성인식기 활용성은 더 높아질 것으로 예측한다.

본고에서는 이러한 음성인식기의 동작 및 내부 알고리즘 등에 대해서 알아 보았고, 국내외 개발사례 등을 살펴 보았다. 음성인식기의 처리방식 및 계산과정은 오랜 연구 끝에 대체로 정형화 되어 왔다. 그러므로 내장형 음성인식기를 요구하는 분야에서는 전력소모 및 가격 측면에서 전용 하드웨어를 개발하여, 시스템의 주 프로세서의 로드를 줄여 주거나 음성인식 결과 지연시간을 줄여주는 성능개선 효과를 기대할 수 있을 것이다.

용어해설

HMM(Hidden Markov Model) 은닉 마르코프 모델의 약자로써, 외부로 관찰되는 관측값의 조합으로 특정 상태열의 조합을 예측하는 확률 모델

GMM(Gaussian Mixture Model) 다수의 가우시안 확률밀도 함수의 조합으로 구성된 함수로써, 자연계에 발생하는 확률 패턴을 기술하기에 적합한 패턴 구분 모델

Viterbi Search 비터비 탐색 알고리즘은 수신되거나 관찰된 신호값의 전후 관계가 확률적으로 미리 기술되어 있어, 수신된 신호 및 관찰된 신호의 정확성 및 오류를 추출하는 알고리즘

약어 정리

HTK	Hidden Markov Model Tool Kit
FPGA	Field-Programmable Gate Array
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
MFCC	Mel-Frequency Cepstrum Coefficient
RTF	Real Time Factor
SoC	System On a Chip
WER	Word Error Rate

참고문헌

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *IEEE*, vol. 77, no. 2, 1989, pp. 257-285.

[2] Y. Choi, K. You, J. Choi, and W. Sung, "A real-time FPGA-based 20,000-word speech recognizer with optimized DRAM access," *IEEE Trans. Circuits Syst. Part I: Regular Papers*, vol. 57, no. 8, Aug. 2010, pp. 2119-2131.

[3] Guangji He et al, "A 40nm 144mW VLSI Processor for Real-Time 60-kWord Continuous Speech Recognition," *IEEE Trans. Circuit Syst. Part I: Regular Papers*, vol. 59, no. 8, Aug. 2012.

[4] CONEXANT, http://www2.conexant.com/Product/Audio/avdspcodecsoc/Pages/default.aspx

[5] sensory, http://www.sensoryinc.com/products/NLP-5x.html

[6] SPANSION, http://www.spansion.com/Applications/VoiceRecognition/Pages/VoiceRecognition.aspx

[7] vr-zone, http://vr-zone.com/articles/meet-intels-hardware-level-sirikiller/60323.html?utm_source=rss