음성 인식에서 가상 스튜디오 기술을 이용한 잡음 제거 방법*

김동현, 유근창, 임준수, 백세인, 이용규
동국대학교 컴퓨터공학과-서울
e-mail: teamammonitedgu@gmail.com

Noise Reduction in Speech Recognition Using Virtual Studio Technology

Dong Hyun Kim, Keun Chang Yoo, Jun Su Lim, Se In Baek, Yong Kyu Lee
Department of Computer Science and Engineering, Dongguk University-Seoul

요 약

최근 음성 인식 기술의 발전으로 음성 인식에 관한 연구가 활발히 진행되고 있다. 음성 인식 기술 중에서도 외부의 잡음을 제거하여 음성 인식의 정확도를 높이는 연구의 필요성이 대두되고 있다. 본 논문에서는 음성 인식에서 가상 스튜디오 기술을 사용하여 잡음을 제거하는 방법을 제안한다. 음성 인식의 전처리 단계에서 잡음 소거 기능을 가진 VST 플러그 인을 사용하여 외부의 잡음을 제거한다. 제안한 방법을 통해 음성인식의 전처리 과정에서 정제되지 않은 음성 데이터로 인해 발생하는 오류를 방지하고 음성 인식의 인식률을 높일 것으로 기대한다.

1. 서론

최근 컴퓨터의 디지털 정보 및 신호 처리 기술이 발전하여 컴퓨터와 인간의 상호 작용을 위한 매개체로써 음성의 역할이 중요해졌다. 이에 따라 음성을 하나의 정보 통신 수단으로 활용하려는 음성 인식 기술에 대한 연구가 활발히 진행되고 있다[1].

초기의 음성 인식 기술은 컴퓨터가 사람 음성을 듣고 저장된 음성 중 입력된 음성과 일치하는 것을 찾아 해당 음성의 유형과 특징을 분석해서 음성을 인식했다. 이러한 음성 인식에서는 음성 데이터의 특징을 확률로 계산해서 나타내는 히든 마커브 모델이 사용되었다. 히든 마커브 모델은 뇌 과학과 컴퓨터과학의 공동 성과물인 인공 신경망(Artificial Neural Network)기술과 결합하면서 현재의 음성 인식 기술의 토대가 되었다. 그러나 음성 인식의 전처리 과정에서 외부의 잡음이 유입되면서 정확한 음성이 인식하지 못한다는 한계가 있다[2].

본 논문에서는 음성 인식에서 가상 스튜디오 기술을 이용한 잡음 제거 방법을 제안한다. 일반적으로 음성이 음성인식 장치에 인식되는 과정에서 외부의 잡음으로 인해 음성 데이터의 품질이 저하된다. 음성 데이터 품질의 저하는 사용자가 원하는 바를 기계가 이행하지 못하는 주요한 문제를 야기할 수 있다. 본 논문에서는 정확한 음성 데이터 추출을 위해 가상 스튜디오 기술(VST, Virtual Studio Technology)의 잡음 제거(Noise Reduction) 플러그 인을 사용하여 외부의 잡음을 제거한다.

음성 인식에서 외부의 잡음을 제거하기 위해 잡음 소거(Noise Canceling) 기술이 사용된다. 잡음 소거 기술은 마이크로폰을 통해 유입된 주변의 잡음을 상쇄해 제거하는 기술이다. 이러한 기술은 유입된 잡음에 대해 역 위상을 갖는 음파를 발생시킨 후, 두 파형을 합성하여 역 위상의 파형을 제거하는 상쇄 간섭(destructive interference) 원리를 사용한다. 일반적으로 음성 인식에서 아날로그 하드웨어가 이러한 원리로 외부의 잡음을 차단한다. 그러나 VST 의 잡음 제거 플러그 인은 아날로그 하드웨어 없이 소프트웨어에서 동일한 원리로 잡음을 제거한다.

본 논문에서 제안하는 시스템을 통해 고가의 아날로그 장비를 구매하지 않고, 소프트웨어 상에서 외부의 잡음을 제거할 수 있다.

2. 관련 연구

2.1. 음성 인식 시스템

기계가 사람의 음성을 듣고 이해하는 데에는 많은 과정들이 필요하다. 음성 인식 시스템[3]은 크게 음성 인식(recognition), 음성에 대한 분석(analysis), 음성 이해(understanding) 세 단계로 나누어진다.

음성 인식 단계에서는 인간의 아날로그 음성 데이터를 디지털 표본화(Digital sampling)하여 디지털 형태

* 본 연구는 과학 기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업의 연구결과로 수행되었음(2016-0-00017)

로 변환(Analog-to-Digital, ADC)한다. 이 단계는 인간의 아날로그 음성 데이터를 디지털로 변환해 컴퓨터가 이해하기 쉬운 형태로 변환하는 단계이다[4].

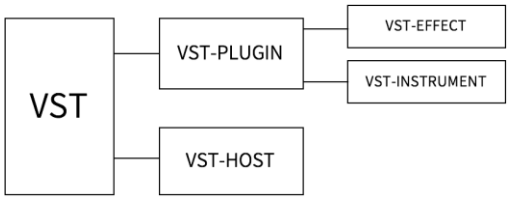
음성 분석 및 이해 단계는 음성의 발성 및 패턴을 파악하여 음성을 의미 있는 단어로 변환하는 단계이다. 이 단계에서는 전처리 된 음성 데이터들을 토대로 파형을 비교하여 적절한 단어로 변환한다.

마지막으로 음성 이해 단계에서는 변환된 단어의 의미를 파악하고 문자 형태로 출력하는 과정을 진행한다. 그러나 음성 이해 단계의 결과는 음성 데이터 품질에 따라 천차만별이다. 특히 음성 분석 단계에서 잡음이 유입된 음성이 제대로 인식되지 않아 다른 의미의 단어로 분류될 가능성이 있다.

본 논문에서는 정확한 음성 데이터 추출을 위해 VST(Virtual Studio Technology) 호스트를 이용하여 잡음을 제거하는 방법을 제안한다.

2.2. VST(Virtual Studio Technology)

(그림 1) 은 가상 스튜디오 기술(VST, Virtual Studio Technology)에서 소프트웨어의 분류를 나타낸 것이다. VST 는 디지털 신호 처리를 사용하여 기존 녹음 스튜디오의 아날로그 하드웨어들을 소프트웨어로 시뮬레이션 하는 기술이다. VST 를 사용하는 소프트웨어는 VST 플러그 인과 VST 호스트로 분류할 수 있다. VST 플러그 인의 종류로는 VST 효과(Effect) 플러그 인과 VST 악기(Instrument) 플러그 인이 있다. VST 호스트(Host)는 이러한 VST 플러그 인들을 불러오기 위한 플랫폼 소프트웨어(platform S/W)이다[5].



(그림 1) VST 를 사용하는 소프트웨어의 분류

본 논문에서는 VST 호스트를 통해 VST 잡음제거 플러그 인을 사용하였다. VST 잡음 제거 플러그 인은 음성 입력 시 유입되는 잡음을 제거하는 VST 효과 플러그 인의 한 종류이다[5].

VST 잡음 제거 플러그 인의 원리는 잡음 소거(Noise Cancelling) 기술이다. 잡음 소거 기술이란 입력된 음성에 대하여 잡음과 역 위상의 관계인 파형을 발생시킨 후, 입력된 음성과 합성하여 잡음을 줄이는 원리를 이용한다. 이러한 현상을 상쇄 간섭(Destructive Interference)이라고 한다.

본 논문에서는 잡음이 유입된 아날로그 음성 데이터를 전처리 단계에서 VST 를 이용하여 잡음을 제거하는 방법을 제안하고, 제안한 방법을 활용한 음성 인식 시스템을 구현한다.

3. VST 를 이용한 잡음 제거 방법

본 논문에서는 음성 인식에서 가상 스튜디오 기술(VST, Virtual Studio Technology)을 이용하여 잡음을 제거하는 방법을 제안한다. 본 논문에서 제안하는 방법을 적용한 음성 인식 시스템은 입력 모듈, 전처리 모듈, 인식 모듈, 출력 모듈로 구성된다.

3.1. 전체 구성도

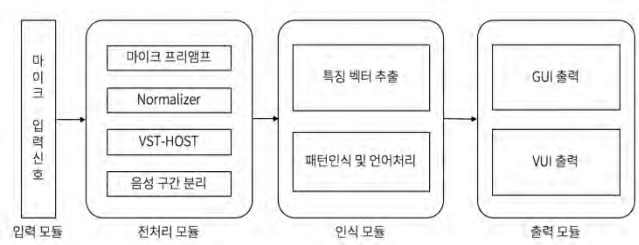
본 논문에서 제안하는 시스템은 입력 모듈, 전처리 모듈, 인식 모듈, 출력 모듈의 네 가지 모듈로 구성되어 있다. 시스템의 전체 구조는 (그림 2)와 같다.

입력 모듈에서는 마이크로폰을 통해 아날로그 음성 신호를 입력 받는다.

전처리 모듈은 음성의 입력을 디지털 표본화 하여 디지털 신호로 변환하고 정규화 한다. 정규화된 신호를 VST 플러그 인을 통해 잡음을 제거하고 음성 구간을 분리한다.

인식 모듈은 전처리가 완료된 음성에 대해 음성 분석 및 음성 이해를 진행한다.

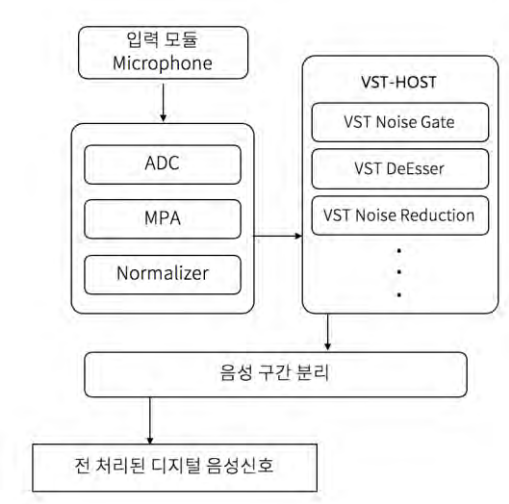
출력 모듈은 최종적으로 처리한 결과를 출력하는 모듈이다.



(그림 2) 음성인식 시스템의 전체 구성도

3.2. 입력 모듈 및 전처리 모듈

(그림 3)는 마이크로폰을 통한 입력 신호의 전처리 과정을 보여준다.



(그림 3)마이크로폰을 통한 입력 신호의 전처리 과정

입력 모듈에서는 마이크로폰을 통해 아날로그 음성 신호를 입력 받으면, 전처리 모듈에 아날로그 음성 신호를 전송한다. 전처리 모듈은 입력 모듈로부터 전송된 아날로그 신호를 디지털 신호로 변환한다. 변환된 디지털신호를 신호의 세기와 상관없이 동일한 수준의 처리를 하기 위해 신호를 정규화하고 증폭시킨다. 디지털 신호로의 변환 과정은 일반적으로 ADC(Analog Digital Converter)가 변환하는 과정과 동일하며, 마이크로폰을 통해 입력된 신호를 증폭하는 것은 MPA (Microphone Pre Amplifier)가 수행하는 과정과 동일하다.

전처리 모듈에서는 임계 값 이상의 의미 있는 음성 신호와 임계 값 미만의 신호인 잡음을 분석하기 위해서, 모든 신호의 다이내믹 레인지(Dynamic Range)를 분석하여 구간을 분리하고 각각 정규화 한다.

3.2.1. 전처리 과정에서의 VST

<표 1>은 마이크로 유입되는 잡음의 종류와 그에 따른 데시벨의 평균 레벨 값을 나타낸 표이다.

<표 1> 잡음 종류에 따른 수음되는 평균 데시벨

타입	수음되는 평균 데시벨(db)
Rustle	- 70 ~ -56
Wind	- 60 ~ -10
Ess	- 30 ~ -10
Mouth Click	- 40 ~ -10
White Noise	- 50 이하
Clip	Random
Hum	Random
Crackle	Random

본 논문에서는 외부의 잡음을 제거하기 위하여 전처리 모듈에서 VST 플러그 인을 이용한다. 이용한 VST 플러그 인은 Noise Gate, De Esser, Noise Reduction 이다.

Noise Gate 플러그 인은 특정 신호가 데시벨(db)를 넘어가지 않으면 차단해서 제거하는 플러그 인이다. 잡음 층(Noise floor)은 평균적으로 -50db 이하의 음성 신호를 말한다. Noise Gate 플러그인은 데시벨의 임계 값(threshold) 이하의 신호를 잡음 층의 음성 신호로 간주하고, 원음에서 제거하는 플러그 인이다.

De-Esser 플러그인은 S 발음의 치찰 음에 대하여 특정 주파수 영역 대의 데시벨을 순간적으로 감소시키는 플러그 인이다. De Esser 플러그인은 마치 컴프레서(Compressor)처럼 동작하며 치찰 음에 의해 중고역대(4Khz~10Khz)의 주파수가 과장되는 것을 방지하는 플러그 인이다.

Noise Reduction 플러그 인은 상쇄 간섭의 원리를 이용하여 잡음을 제거하는 플러그 인이다. Noise Reduction 플러그 인은 잡음 구간과 음성 신호 구간을 분석하여 잡음 구간의 역 위상 관계인 파형을 발생시킨다. 발생된 파형을 원본 파형과 합성시켜 상쇄 간섭의 원리를 통해 잡음을 제거한다.

3.3. 인식 모듈과 출력 모듈

인식 모듈에서는 전처리가 완료된 신호에 대하여 스펙트로그램(Spectrogram)으로 음성 신호의 구간을 디지털 파형의 샘플단위의 특징 벡터로 추출한다. 파형 샘플에서 추출한 특징벡터의 패턴에 대해 데이터베이스에 저장되어 있는 적절한 패턴과 매치하여 언어 처리한다.

처리된 언어를 인식하면, 인식 결과를 출력 모듈로 전송하여 출력 모듈에서는 최종 결과를 음성 유저 인터페이스(VUI) 또는 그래픽 유저 인터페이스(GUI)로 표현한다.

4. 실험 및 구현

본 논문에서 제안하는 방법을 적용한 음성 인식 시스템은 사용자가 말한 문장을 음성이나 문자의 형태로 출력할 수 있다.

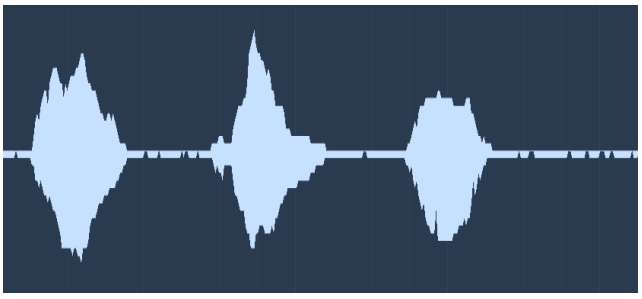
4.1. 실험

본 논문에서는 VST 플러그 인을 통한 잡음 제거 방법의 성능을 확인하기 위한 실험을 진행하였다.

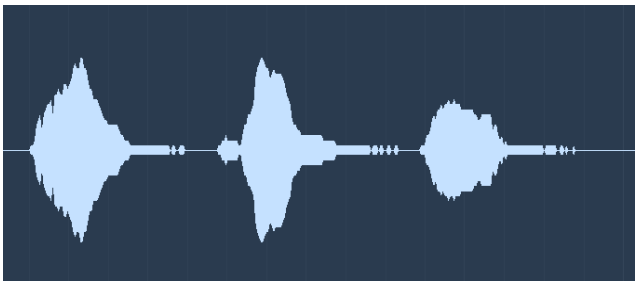
실험에서 사용된 샘플을 녹음하기 위하여 오디오 인터페이스(Audio Interface)로 M-Track 2X2M, 컨덴서 마이크(Condenser microphone)인 Mxl2006 이 사용되었다.

실험이 진행된 환경은 조용한 스튜디오에서 진행하였으며, 기계적 잡음(Mechanical Noise)와 백색 소음(white noise)를 유입하였다.

실험에 사용된 단어는 “꾀꼬리”이다. (그림 4)는 VST를 사용하지 않은 원본 파형을 나타낸다. 음성 구간 사이의 잡음이 끊임없이 유입되어 있는 것을 확인할 수 있었다.



(그림 4) VST 플러그 인을 활용하지 않은 음성 파형



(그림 5) VST 플러그인을 활용한 음성 파형

(그림 4)는 VST 플러그 인을 통하여 잡음을 제거한 파형을 나타낸다. 음성 구간 사이의 잡음이 비교적 제거되었으며, 다이내믹 레인지(Dynamic Range) 또한 정렬되어 있는 것을 확인할 수 있었다.

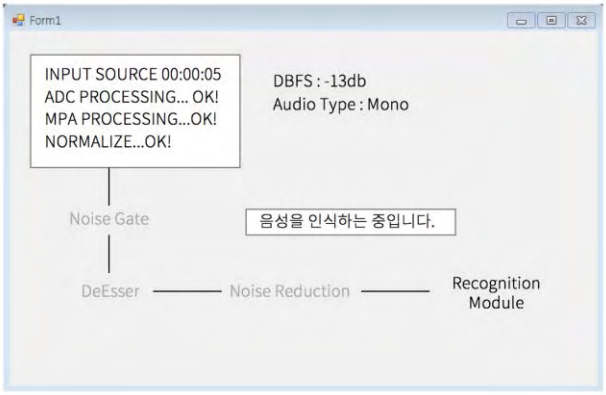
4.2. 구현 결과

본 논문에서 구현한 음성 인식 시스템은 MFC(Microsoft Founded Classes)를 기반으로 하고 있다. C++ 언어를 사용하였으며, 서버는 웹 서버인 FIREBASE 를 사용하여 구현하였다. VST 가 동작할 수 있도록 내부적으로는 VST-HOST 로 구현하였다.

사용한 라이브러리로는 언어처리를 위해 Microsoft 가 제공하는 Microsoft Speech C++ API 를 사용하였고, MFC 의 클래스 라이브러리를 사용하였다.

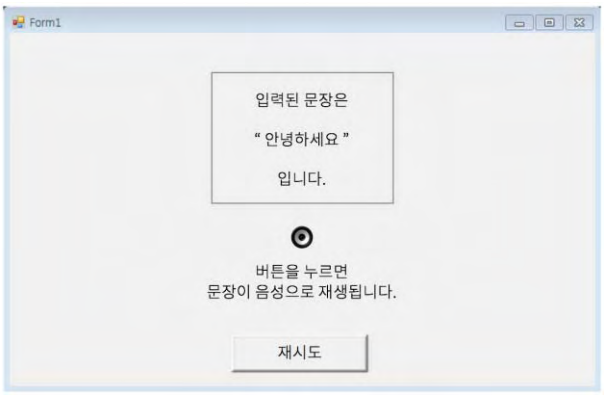
음성 인식 시스템은 초기화면, 음성 입력을 받기 위한 화면, 신호 처리를 보여주는 화면, 결과를 출력하기 위한 화면으로 구성되어 있다.

(그림 6)은 신호를 처리하는 과정을 출력하는 화면이다. 사용자는 초기화면에서 마이크 버튼을 누르고 음성을 입력한다. 입력이 완료되면 다음 화면에서 신호를 처리하는 과정을 출력한다.



(그림 6) 신호 처리 과정을 출력하는 화면

(그림 7)은 음성 인식 결과를 출력하는 화면이다. 음성 인식이 완료되면 인식 결과를 음성과 문자열로 출력한다.



(그림 7) 음성 인식 결과를 출력하는 화면

5. 결론

본 논문에서는 음성 인식에서 가상 스튜디오 기술을 이용한 잡음제거 방법을 제안하였다. VST 플러그 인을 이용한 잡음 제거 방법을 적용한 음성 인식 시스템을 구현하였다.

기존의 잡음 제거 방법은 고가의 아날로그 하드웨어 장비를 통해서 잡음을 제거한다. 또한 Noise Gate 의 특성을 하드웨어로 구현하여 소음일 가능성이 높은 음성신호를 제거하는 방법도 사용되고 있다. 본 논문에서는 별도의 하드웨어 장비 없이 VST 플러그 인을 통한 잡음 제거 방법을 제안하고, 이를 적용한 음성 인식 시스템을 구현하였다. 실험 결과, VST 플러그 인을 이용하여 외부의 잡음 처리가 된 것을 확인할 수 있었다.

향후 음성 인식의 정확도 평가를 수행하고, 화자 검증이 가능한 음성 인식 시스템으로 확장할 예정이다.

참고문헌

[1] 최재승, "음성특징벡터 및 정규화 인식방법을 이용한 화자종속 음성인식", 한국정보기술학회 논문지, Vol. 10, No. 5, pp. 61-66, 2012.5.

[2] 이혜민, 김형순, "HMM 기반 한국어 음성합성에서의 화자적응 방식 성능비교 및 지속시간 모델 개선", 한국음성학회, 말소리와 음성과학, Vol. 4, No. 3, pp. 111-117, 2012.9.

[3] 김동헌, "핵심어 추출 및 연속 음성 인식 지원을 위한 다목적 처리 프로세서 설계에 대한 연구", 전남대학교 박사학위논문, 2013.8.

[4] Gyu-Ha Choe, Jeong-Woo Kim, Young Hoon Cho, "An Analysis of ZVS Phase-Shift Full-Bridge Converter`s Small Signal Model according to Digital Sampling Method", 전력전자학회 논문지, Vol. 20, No. 2, pp. 167-174, 2015.4.

[5] 위키피디아, 가상 스튜디오 기술, https://en.wikipedia.org/wiki/Virtual_Studio_Technology

[6] Kuk-Hee Lee, Se-Kyo Chung, Byeong-Geuk Kang, Yong Oh Choi, Jae-Sun Won, Hee-Seung Kim, "Design and Implementation of an Active EMI Filter for Common-Mode Noise Reduction", Journal of Power Electronics, Vol. 16, No. 3, pp. 1236-1243, 2016.5.