멀티미디어시스템 REPORT

자동 음성 복원 기술 연구

2020 년 06 월 26 일

제출: 32194747 최지윤

목 차

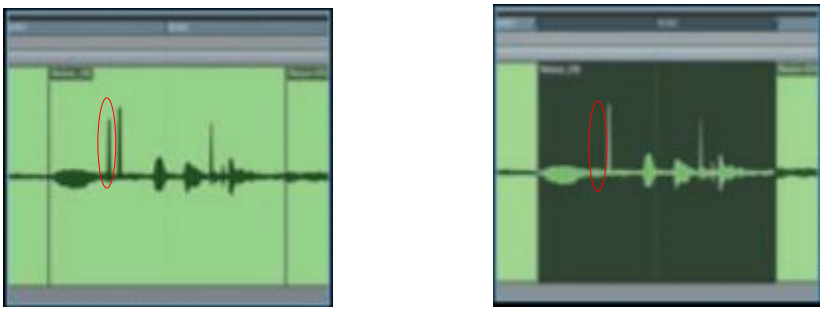
Ⅰ. 연구 개요

- 기술 배경 및 필요성

음성 데이터를 저장하기 위한 기술이 발전하면서 데이터 손실이 따라오는 아날로그형 음성을 디지털 형태의 음성으로 변환하기 위한 기술의 필요성이 대두되었다. 이에 따라 아날로그를 디지털로 변환하는 과정에서 생겨난 왜곡이나 잡음에 대해 원래의 음성 데이터로 복원하려는 작업이 매우 중요해졌고, 원본의 형태는 최대로 유지하면서 깨끗하고 높은 음질을 제공할 수 있는 기술의 개발이 필요하게 되었다. 즉, 아날로그 데이터 자체의 물리적 변형, 변환 과정에서 생기는 왜곡으로 인해 다양한 형태의 잡음이 발생하는데 이러한 잡음을 처리하는 기술의 발전으로 우리는 음성 데이터를 보다 효과적으로 저장 및 보관 할 수 있게된다.

- 기존 기술의 현황, 문제점 및 개선 방안, 기대효과

기존의 오디오 신호를 복원하는 프로그램은 사용자가 지정한 잡음만을 처리할 수 있는 반자동 형태였다. 대표적으로 iZotope RX라는 프로그램이 있다. iZotope RX에서는 시각화된 음성 신호를 제공하여 사용자가 잡음의 위치와 형태를 파악한 뒤 [그림 1]과 같이 클릭이나 드래그 등을 통해 개별적으로 잡음신호를 처리할 수 있다.



[그림 1] iZotope RX 프로그램 실행

하지만 모든 음성을 직접 청취하면서 훼손되거나 왜곡된 부분을 찾아내어 일일이 수정해야하기 때문에 효율성이 떨어지고 많은 양의 데이터 처리에 있어 불편하다는 문제점이 있다. 또한, 기존의 잡음 제거 방법을 사용하여 잡음을 축약시키거나 제거할 경우, 필요한 신호성분도 같이 왜곡되거나 제거되는 경우가 발생하기도 한다. 특히 각종 돌발 잡음의 경우, 확률적 특성이나 상관함수가 없어 예측 불가능하기 때문에 이러한 잡음을 처리하는 데에는 큰 어려움이 있었다.

최근 개발된 자동 형태의 복원 기술은 기존의 반자동 형태의 기술의 최대 단점인 효율성을 보완해준다. 자동 형태의 복원 기술은 자동으로 잡음의 위치를 검출하고 특

성에 따라 신호를 분류하여 적합한 복원 조치를 취하는 기술을 뜻한다. 이는 사용자가 수많은 음성 데이터를 직접 들어보지 않아도 정확하고 빠르게 훼손 부분을 찾아 처리하기 때문에 시간적, 비용적 측면에서 이득을 보면서 복원 효과 또한 뛰어나다는 평가를 받고 있다.

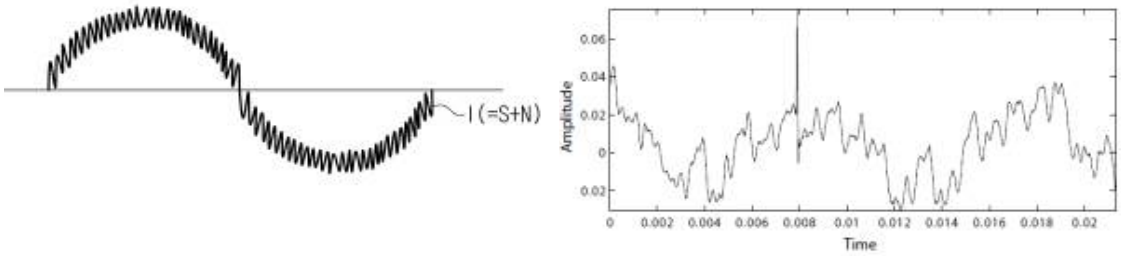
Ⅱ. 연구 내용

- 음성 상태 분류

높은 음질의 음성 데이터를 저장하기 위해서는 음성이 훼손되거나 왜곡된 부분이 있는지, 음성의 상태는 어떠한지 등 음성 신호의 특성을 파악해야 한다. 음성 상태의 분류에 따라 복원이 가능할지 결정하고 적절한 복원 방법을 사용할 수 있다.
음성의 상태는 정상, 음량 문제, 음성 왜곡, 음성 소실 등 5가지로 구분할 수 있다. 정상 신호는 말 그대로 음성이 일절 훼손되지 않은 깔끔한 음질을 뜻한다. 음량이 일정하지 않거나 적당하지 않아 문제가 되는 신호는 음량 문제 카테고리에 속하고, 의도치 않은 잡음이 들어갔거나 음성 자체의 왜곡으로 오염된 신호는 각각 잡음과 음성 왜곡으로 구분한다. 마지막으로 음성 소실은 손상이 심하거나 손실이 생긴 음성 신호를 의미한다. 정상 신호의 경우에는 음질 향상을 위한 어떠한 대처도 필요하지 않겠지만 음량 문제, 잡음, 음성 왜곡이 있는 신호의 경우에는 훼손이 생긴 부분을 추정하고 복원하는 과정이 필요하다. 반면에 음성 소실이 있는 신호는 훼손 정도가 심하기 때문에 현재로선 복원하는 것이 크게 의미가 없다.

- 잡음의 종류

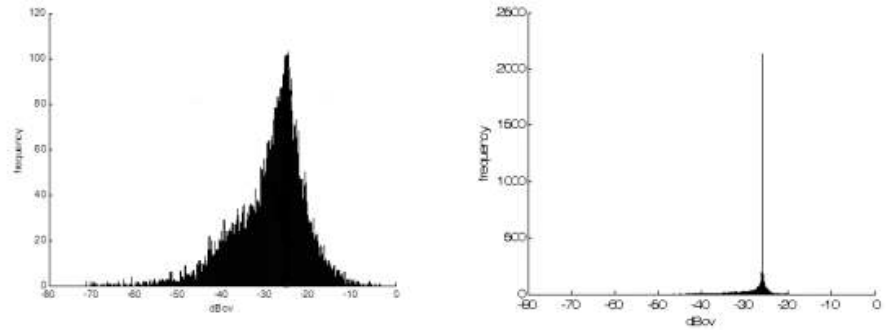
잡음은 크게 두가지로 구분한다. 정상 잡음과 돌발 잡음이다. 정상 잡음은 확률적 특성을 가지고 변화가 별로 나타나지 없는 신호를 뜻하고, 돌발 잡음은 반대로 외부적, 물리적인 작용에 의해서 예기치 않게 발생하는 잡음을 뜻한다. 정상잡음의 예로는 전기적 특성이나 자기장 영향으로 펼쳐져 있는 잡음이 있고, 주로 음성 신호 전체에 골고루 분포되어 있다. 돌발 잡음은 특정 구간에 일시적으로 나타나며 클릭잡음이나 키보드 치는 소리, 문 닫는 소리와 같은 생활 소음 등 연속적이지 않은 잡음이 포함된다. 두 잡음의 특성은 명확하게 구분되기 때문에 그 처리 및 복원 방법 또한 다르다.



[그림 2] 정상 잡음이 섞인 신호 / 돌발 잡음이 섞인 신호

- 음량 복원

음량은 전체 구간이 아닌 음성이 존재하는 구간 내에서 측정한다. 이를 유효 음압 수준이라고 한다. 음성 신호의 유효 음압 수준은 샘플링한 음성의 포락선이 임계값을 넘게되면 그때마다 수를 누적하는 방식으로 측정한다. 따라서 유효 음압이 일반적 음압보다 작으면 '음량 작음', 크면 '음량 큼'이라고 판단한다. 음량이 작은 신호 부분과 음량이 큰 신호 부분은 각각의 이득값을 곱해주는 식으로 일반적 음압으로 일정하게 음량을 조절한다.



[그림 3] 유효 음압 수준이 일정하지 않은 음성 / 음량 조절 후 음성

- 정상 잡음 제거

잡음을 복원하기 이전에는 기본적으로 음량을 조절하는 과정이 동반 된다. 하지만 음성신호에는 정상 잡음이 더해져 있기 때문에 음량을 조절하게 되면 정상 잡음의 크기까지도 변하게 되는 문제가 발생한다. 이러한 경우, 정상 잡음을 예측하고 음질 향상기를 사용하여 효과적으로 복원할 수 있어야 한다.

정상잡음은 음성보다 상대적으로 느리게 변화하고 일정한 분포를 갖는다는 특성이 있다. 이 특성을 이용하면 잡음 성분을 추정하여 제거할 수 있다. 우선적으로 잡음

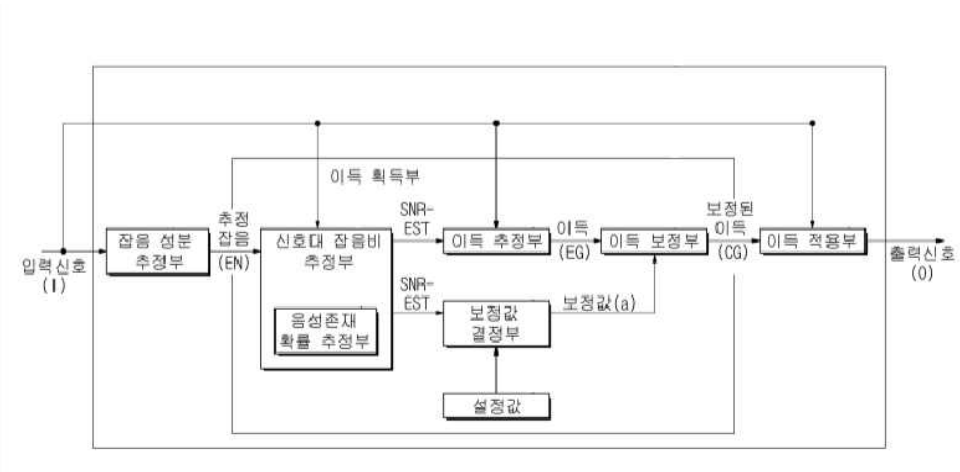
추정부에서 음성이 존재하지 않는 구간의 신호를 측정한 후 잡음의 파워 스펙트럼을 추정한다. 이때 도출된 잡음의 스펙트럼은 신호 대 잡음비 추정부에 사용된다. 신호 대 잡음비는 잡음 세기에 대한 신호 세기를 나타내는 비율로, 음성 신호에 정상 잡음이 얼마만큼 포함되어 있는지 파악할 수 있다. 기본적인 신호 대 잡음비(SNR)의 수학적 정의는 다음과 같다.

$$SNR = c\log(\frac{S^2}{N^2}) \quad \cdots\cdots(1)$$

여기서 N은 잡음, S는 잡음이 섞이지 않은 원신호를 의미하며 c는 선택적인 상수이다. 이렇게 계산된 신호 대 잡음비는 이득 추정부에서 이득을 추정하고, 추정된 이득을 보정하기 위한 보정값을 설정하는 데에 사용된다. 보정값은 잡음 신호의 특성에 따라 0과 1 사이에서 결정되는데, 보정값이 1에 가까울수록 잡음이 제거된 신호가 많아지고 0에 가까울수록 잡음이 제거된 신호가 적어진다.

보정이 완료된 이득을 본래 입력 신호에 곱해주면 입력 음성 신호에서 잡음 스펙트럼이 빠지면서 잡음이 제거된 음성 신호를 추출할 수 있게 된다. 이때, 큰 보정값과 이득값을 사용하게 되면 잡음이 제거되는 효과는 증진되겠지만 본래 음성까지 훼손될 수 있는 위험이 있기 때문에 적절한 보정값과 이득값 설정이 매우 중요하다.

음성 존재 확률 추정부에서는 음성이 존재하는지 부재하는지를 고려하여 이득추정값을 수정하는 역할을 한다. 최소값 제어 재귀 평균 (MCRA) 알고리즘을 사용하여 변화가 큰 잡음환경에서의 성능을 보완해준다.



[그림 2] 정상 잡음 음질 향상기 블록도

- 돌발 잡음 제거

돌발 잡음은 주로 짧은 시간에 큰 진폭을 가지면서 잡음 시작 부분에서 큰 에너지로 나타났다가 사라지는 특성을 갖는다. 돌발 잡음의 크기가 음성 신호 보다 진폭이 훨씬 클 경우에는 진폭만을 이용해서 잡음의 위치를 쉽게 검출할 수 있겠지만, 대부분의 경우와 같이 두 신호의 진폭이 비슷하거나 돌발 잡음의 크기가 더 작을 때에는 잡음의 위치와 그 크기를 예측하기는 쉽지 않다. 따라서 고대역에서 높은 에너지를 갖는다는 돌발 잡음의 특성과 에너지가 급격하게 사라져 변화가 크다는 특성을 이용하여 잡음을 검출 하는 방법을 채택하고 있다.

2차 미분 계수는 흔히 기울기 값을 구하는 데에 사용된다. 변화가 큰 부분일수록 큰 값과 큰 에너지를 가지기 때문에 이를 입력 음성 신호의 단구간 에너지에 적용해 보면 신호가 급하게 변하는 위치를 검출할 수 있고 이 위치는 대게 돌발 잡음의 위치일 것이다. 입력 호를 $x[n]$이라고 하고, 2차 미분 계수 $z[n]$을 표현하면 다음과 같다.

$$z[n] = D^2 x[n] = x[n-1] - 2x[n] + x[n+1] \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(2)$$

이때, 2차 미분 계수의 에너지는 다음과 같이 구할 수 있다.

$$w[n] = \left[\frac{1}{N+1} \sum_{j \to N/2}^{N/2} z^2[n+j] \right]^{\frac{1}{2}} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(3)$$

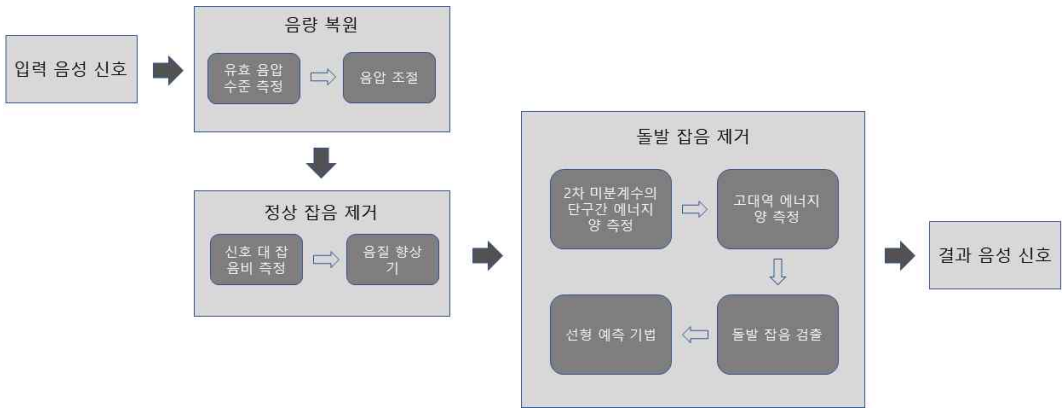
구해진 2차 미분 계수와 단구간 에너지를 이용하여 배경 신호를 추정하고 배경 신호보다 값이 급하게 커지는 부분을 찾아내야 한다. 그러기 위해선 미디언 필터를 사용해야 한다. 이 필터는 데이터 값의 중앙값을 결정해주는 필터로, 배경 신호와 비교하여 급격한 변화가 생기는 부분이 어디인지 파악할 수 있게 도와준다.
또, 음성 신호는 주파수의 저대역에 에너지가 집중되어 있기 때문에 고대역에 있는 에너지를 추출해보면 돌발 잡음의 위치를 예측할 수 있다. 마찬가지로 기준이 되는 배경 신호를 설정하고 미디언 필터를 사용해서 배경 신호에서 보다 큰 값을 갖는 부분을 찾아낸다.

에너지 변화만 분석하면 돌발 잡음의 시작 위치를 파악하는 것은 그리 어렵지 않지만 어느 위치까지 돌발 잡음으로 인식하여 제거햐야 하는 지가 가장 중요하다. 돌발 잡음 구간을 잘못 잡게 되면 본래의 음성 신호까지 훼손될 수 있기 때문이다. 따라서 최종적인 돌발 잡은 구간은 고대역의 에너지가 큰 구간 안에서의 2차 미분 계수와 그 에너지 또한 큰 부분만으로 한정한다. 결과적으로는, 에너지가 급하게 커지기 직전, 즉 잡음이 등장하기 이전의 에너지값과 같거나 비슷해지는 위치까지가 돌발 잡음 구간이 된다.

돌발 잡음 구간을 검출했으면 잡음을 제거하고, 제거한 부분이 어색하지 않게 복원하는 과정이 필요하다. 돌발 잡음 구간에서 인접한 음성 신호를 선형 예측하여 대체하는 방식으로 복원할 수 있다. 이때, 음성의 주기적 특성인 피치는 따로 구해 더해주어야 한다. 선형 예측 필터의 잔차 신호에 장구간 예측 기법을 사용하여 피치의 정보를 얻어내 원래 신호에 저장한 후 잡음이 제거된 신호에 더해주는 방식으로 피치를 합성한다. 이와 같은 방법으로 보다 효과적으로 향상된 음질을 도출할 수 있다.

- 정리 및 결론

복잡한 수식들은 뒤로하고 내가 이해한 자동 음성 복원의 알고리즘을 정리해 다음과 같이 그려보았다.



[그림 3] 자동 음성 복원 알고리즘

개선된 음질을 제공하기 위한 자동 음성 복원은 음량 조절, 정상 잡음 제거, 돌발 잡음 제거 순서로 이루어 진다. 음량은 음성이 존재하는 구간에서 유효 음압 수준을 구하고, 적당한 음압 수준 값과 비교해 일정하게 조절하면서 복원시킬 수 있고, 이후에 정상 잡음을 처리하기 위해서는 잡음의 스펙트럼을 추적해서 신호 대 잡음비를 측정한 후 음질 향상기를 이용한다. 반면에 돌발 잡음은 고주파에서 높은 에너지를 갖고 있음을 이용하여 2차 미분 계수의 단구간 에너지와 검출된 고대역 에너지 양에 따라 선형예측기법을 통하여 처리한다. 이와 같은 순서로 복원해야하는 이유는 음량이 조절되면서 정상 잡음이 강해지거나 희미해질 수 있어 잠음 제거 후 음량 조절은 무의미하고, 정상 잡음이 제거된 후 돌발 잡음을 검출할 때 돌발 잡음 검출이 용이해지고 결과값이 더 정확해 복원 효과가 뛰어나기 때문이다.

- 응용 분야

1) 음성인식을 위한 음질 개선

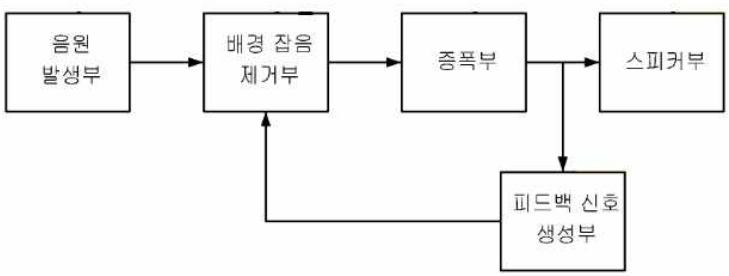
음성인식 기술은 컴퓨터가 음성 신호를 인식하고 분석하여 단어나 문장으로 변환하는 과정으로 STT (Speech-to-Text)라고도 한다. 최근 딥러닝 기반의 음성인식 기술로 인간과 컴퓨터가 자연스럽게 대화하는 것도 가능할 정도로 발전했다.
이때, 음성인식 기술의 원리를 알아보면 다음과 같다.

[그림 4] 음성인식 과정

[그림 4]에서 음성을 인식하기 위한 전처리부에 인식 구간을 추출하여 잡음을 처리하는 과정이 들어가 있음을 알 수 있다. 컴퓨터가 사람의 언어를 보다 적확하게 이해하기 위해서는 음성이 명확하게 들려야 한다. 하지만 마이크로 들어가는 음성에는 어쩔 수 없이 마이크 잡음이나 주변 소음이 포함된다. 이는 음성인식 과정에 걸림돌이 되기 때문에 전처리부에서 우선적으로 처리하게 된다. 따라서 음성 복원 기술을 사용하여 적절히 음량을 조절하고, 마이크 잡음인 정상 잡음과 주변 소음인 돌발 잡음을 제거하는 과정을 거치기 때문에 깨끗한 음질과 명확한 발성으로 컴퓨터가 정확하게 인식할 수 있도록 도와준다.

2) 잡음 제거 스피커

양질의 오디오 출력을 위해서 스피커에 음질 향상기를 사용하는 스피커가 개발되었다.



[그림 5] 잡음 제거 스피커 구조

[그림 5]와 같은 배경 잡음 제거부에서는 음량을 일정화하고 마이크 잡음 등 블필요한 잡음을 제거한다. 이렇게 잡음이 제거된 신호는 증폭부를 통해 증폭되어 스피커로 출력된다. 음성 신호를 따로 잡음 제거 및 음성 복원 프로그램에 넣어 결과를 가져오지 않아도 스피커 내에서 향상된 음질의 음성을 제공한다는 점에서 유용하다.

Ⅲ. 제안 사항

- 기술 제안

앞서 음원 소실이 있는 음성 신호는 복원이 현재로선 무의미하다고 소개하였다. 이에 소실된 음성 신호를 어떻게 복원할 수 있을지 고민해 보았다. 하지만 입력 신호가 존재해야 깨끗한 음질로 복원이 가능할 텐데, 소실되어 입력 신호가 없는 경우에는 일반적인 복원 방법으론 개선 불가능하다고 판단하였다. 그렇다면 음질 복원이 아니라 소실된 부분의 음성을 재합성하여 대체하면 어떨까? 소실이 있는 부분 외의 정상 신호의 특성을 반영하여 음성을 만들어낸 뒤 소실 구간에 합성하는 것이다.

음성 합성은 주로 인공지능 스피커와 같은 음성 대화 기술에서 쓰이는 기술로, 모델이 되는 음성 데이터를 분석한 후 일정한 음성 단위로 분할하고 필요에 따라 음성 단위를 합쳐 소리를 만들어 낸다. 이러한 음성 합성 기술을 이용한다면, 우선 소실된 음성 신호를 예측 및 재구성하고 소실 되지 않은 정상 신호의 음성 데이터 분석을 통해 합성하는 방식으로 복원할 수 있을 것 같다. 물론 정상 신호의 데이터가 소리를 만들어 낼 정도로 충분하지 않거나 소실된 음성의 구간이 넓을 때에는 복원에 어려움이 있겠지만 음성 합성 기술을 적용해 기존엔 복원 불가능으로 분류되었던 소실 음성을 어느정도 개선 시킬 수 있다는 것에 의의가 있는 것 같다.

이번엔 웅성웅성하는 주변의 소리가 사람의 음성 전체 신호에 퍼져 있는 경우에 대한 복원 방법을 생각해 보았다. 이러한 소음은 변화가 별로 없거나 일정한 분포를 갖지도 않고, 그렇다고 짧은 시간에 큰 에너지로 등장하지 않기 때문에 정상 잡음이나 돌발 잡음에 해당되지 않는다고 판단하였다. 따라서, 음성 신호를 인식하고 분석한 다음 명확하게 들리는 유효 음성 언어만을 남기고 나머지 배경음은 제거하여 복원하고자 한다. 이를 위해선 음성 인식 기술과 같은 맥락으로, 음성 신호 속 언어를 추출해서 언어 모델을 사용한 인식 알고리즘을 통해 언어를 처리하는 기술이 내장되어야 한다. 이렇게 언어처리를 이용하여 컴퓨터가 언어, 단어 및 문장으로 인정한 신호인 ‘유효 음성 언어 신호’가 추출되면 나머지 신호는 제거하거나 음성 합성으로 재구성할 수 있을 것 같다.

이 기술은 어디에 어떻게 활용될 수 있을까? 가장 먼저 콘서트장이 떠올랐다. 콘서트장의 특성상 큰 음향과 함께 들리는 다운 비트 잡음, 주변 사람들의 소음, 마이크 잡음 등은 가수의 원 목소리와 깨끗한 반주를 들으며 감상하고자 하는 사람들에게 큰 방해요소가 될 수도 있다. 따라서 음향 조절, 주변 잡음 제거, 기계 잡음 제거 등의 기술을 넣어 콘서트장 전용 이어폰을 개발한다면 깨끗한 음향과 음질, 선명하게 들리는 가수의 목소리 등을 감상할 수 있을 것이다. 여기에 추가적으로 반사(울림) 음향이나 흡음(먹힘) 음향을 보정하는 기술을 연구하여 적용하면 뛰어난 효과를 볼 것 같다.

- 멀티미디어 후기

갑작스러운 코로나 사태로 대면 강의가 불가능해져 온라인강의를 수강하게 되면서 많은 혼란이 생겼지만 그 중 멀티미디어 수업은 대면 강의와 차이가 없다고 느낄 정도로 깔끔했다. 수업 내용 또한 흥미롭고 필기하는 재미에 빠져 공부가 하기 싫을 때 찾아서 공부하는 유일한 과목이 멀티미디어였다. 기말과제 또한 주제 선정과 연구 내용 학습에 많은 시간과 노력이 들어가는 만큼 학습 효과도 뛰어나고 하나의 레포트가 완성되었기 때문에 개인적으로 도움이 많이 되었다는 느낌을 받았다. 다만, 우려되는 점은 고학번의 경우 각종 프로젝트와 레포트를 다루어본 경험이 많아 유리할 수 있다는 것이다. 하지만 나에게 이번 연구 레포트는 평가를 넘어서 나의 지식과 경험을 쌓을 수 있는 좋은 기회였음이 분명하다.

이번 과제의 연구 주제로 자동음성복원 기술을 선택한 데에는 온라인 강의의 영향이 있었다. 온라인 강의를 듣다보니 간혹 잡음 때문에 교수님의 음성이 잘 들리지 않는 경우가 발생했다. 따라서 이러한 잡음을 없애고 음질을 높일 수 있는 복원 방법에 대해 연구하고 싶은 마음이 생겼다. 주제를 정한 뒤 관련된 연구 내용을 학습하려고 찾아보았는데 쉽게 이해가 가지 않았다. 생소한 기술 용어, 복잡한 기술 내용 때문이었다. 몇 번이고 주제를 바꿔볼까 고민하였지만 포기하지 않고 관련 연구 보고서, 논문, 그리고 각종 사이트에 기재된 기술 설명 등을 열심히 정독하고 반복하여 읽었더니 어느 순간 머릿속에 틀이 잡히기 시작했고 레포트를 막힘없이 작성할 수 있었다. 많은 시간과 노력을 쏟았던 만큼 뿌듯함도 크다. 앞으로 기회가 된다면 이미지 복원이나 음성 인식 및 합성 기술에 대해서도 연구해보고 싶다.

Ⅳ. 참고 자료

(1) 훼손된 음성기록물 복원 기술 연구 동향 / 국가기록원
(2) 국가기록원 음성 기록물의 복원과 분석 / 국가기록원
(3) 음질 개선을 위한 돌발잡음 제거와 음성 복원 / 한국정보통신대학원대학교 음성 음향 연구실 등