



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

석사학위논문

# 오디오 스타일변환 기반 음성 변환을 이용한 고령 화자의 음성 데이터 전처리 기술에 관한 연구

A Study for the elderly Speech Data Preprocessing Technique  
with Voice Conversion based on Audio Style Transfer

한국방송통신대학교 대학원

정 보 과 학 과

우 은 주

2019년

# 오디오 스타일변환 기반 음성 변환을 이용한 고령 화자의 음성 데이터 전처리 기술에 관한 연구

A Study for the elderly Speech Data Preprocessing Technique  
with Voice Conversion based on Audio Style Transfer

지도교수 이 병 래

이 논문을 석사학위 논문으로 제출함


한국방송통신대학교 대학원

정 보 과 학 과

우 은 주

2019년 12월

우 은 주의  
석사학위 논문을 인준함

심사위원장 김강현 

심사위원 정재화 

심사위원 이병래 

한국방송통신대학교 대학원

2019년 12월

# 논 문 요 약

오디오 스타일변환 기반

음성 변환을 이용한

고령 화자의 음성 데이터

전처리 기술에 관한 연구

우 은 주  
한국방송통신대학교 대학원

정 보 과 학 과

(지도교수 이병래)

CUI (Conversational User Interface)란 오랫동안 진행된 인간과 기계 상호작용의 결과로 사람들의 자연스러운 대화를 모방하여 컴퓨터와 상호작용하는 새로운 패러다임이다. 이는 특히 전자기기 작동에 익숙지 않아 쉽게 사용을 할 수 없는 고연령층을 위한 기술이다. 한국은 2018년 65세 이상인 고령 인구의 비율이 14%를 넘어가 본격적인 고령사회로 접어들었다. 이러한 사회 현상에도 불구하고 한국에서는 고령화자 음성의 특성과 그의 개선

및 복원에 관한 연구는 다른 나라에 비해 거의 진행되지 않고 있다. 고령화자 음성은 그 소리가 작고, 발음이 부정확하여 젊은 층의 표준발음을 기준으로 하여 만든 일반적인 음성 인식기에 적합하지 않다. 이러한 고령 화자의 음성 특성을 고려하여 본 연구는 오디오 스타일변환을 이용해 고령 화자의 음성 데이터 전처리를 통해 거동이 불편한 독거노인, 노인성 음성 질환을 앓고 있는 환자 등이 사용하는 음성인식 시스템의 인식 성능을 높여 사용의 편리함을 도모하고 단순한 잡음 감쇄, 정규화 등의 전처리 효과만으로는 원본 데이터에 없는 새로운 데이터를 새로 생성할 수는 없는 기존 음성 기술의 단점을 보완하기 위해 진행하였다.

본 연구에서는 2005년 국립국어원에서 보급한 서울말 낭독체 발화말뭉치를 이용하여 실험을 진행하였다. 이 데이터 중 음성의 질이 가장 떨어지는 고령 화자를 성별로 각각 한 명씩 선정하였다. 이렇게 선정한 68세 여성, 71세 남성 화자의 음성 데이터 33개를 음성 변환을 수행하였다. 오디오 스타일변환에 사용한 스타일 음성은 일본 나고야 대학에서 개발한 SPRocket의 결과물이다. 이를 사용하여 원시 화자를 고령화자, 목표 화자를 20대의 젊은 화자로 하여 음성 변환시킨 결과물을 사용하였다.

본 연구에서 제안한 딥러닝 알고리즘은 2D Random CNN을 두 번 훈련하는 것으로 구성된 SeniorCNN이다. 노인 음성에 SPRocket으로 음성 변환시킨 음성의 스타일을 입히는 용도로 설계한 SeniorCNN을 작동시키기 위해서는 두 개의 스타일 음성이 필요하다. 첫 번째 단계에서는 GMM 기반의 전통적인 음성 통계학적인 기법을 사용한 일반 음성 변환 결과를, 두 번째 단계에서는 보코더 없이 변환한 음성의 음성 변환 결과를 스타일 음성으로 사용하였다.

음성 변환의 결과 분석을 위해 SeniorCNN 훈련을 거친 음성 데이터를 ETRI에서 공개한 음성인식 API에 인식시켜 음성 변환 작업 단계별 단어 오류율을 비교하였고, 음성품질 평가의 객관적인 척도로 사용되는 PESQ와

이를 변환한 MOS 값을 이용하여 SeniorCNN의 결과 음성을 SPRocket의 결과 음성과 원래 음성 데이터와 비교하여 음성인식 성능을 측정하였다.

단어 오류율은 SeniorCNN의 두 단계를 모두 통과시켰을 때 전반적으로 3% 정도 증가하였으나 ETRI 음성인식 결과는 음성 데이터를 전송해 API에서 인식시킬 때마다 항상 일정한 텍스트를 출력하는 것이 아니고 짧은 화자의 음성도 단어 오류율이 10% 정도인 것을 참작하여 결과를 분석해 보니 SeniorCNN 훈련 전 원래 고령 화자의 음성에서 인식하지 못한 단어와 어구들을 SeniorCNN의 훈련을 통과한 음성으로 음성인식 시켰을 때는 정확한 단어와 어구로 인식하기도 한다는 것을 확인할 수 있었다.

변환한 음성별 PESQ와 이를 MOS로 변환한 값을 구해 분석한 결과 SPRocket으로만 음성 변환한 값보다는 이를 스타일 음성으로 사용하여 SeniorCNN으로 훈련한 음성 데이터의 PESQ, MOS 모두 더 큰 값을 나타냈다.

SeniorCNN 훈련결과는 첫 번째 CNN 훈련만 통과시킨 중간 결과 음성이 두 번째 CNN까지 통과시킨 것에 비해서 단어 오류율과 PESQ, MOS 모두 더 우월한 수치를 나타내어 CNN 훈련을 여러 번 반복하는 것이 반드시 좋은 결과를 나타내는 것은 아니라는 것을 확인할 수 있었다.

또한, 본 연구는 CNN 훈련 시 사용하는 활성화 함수에 따라 최종 결과 음성의 품질이 어떻게 달라지는지를 비교 및 분석하였다. 가장 많이 쓰이는 활성화 함수 ReLU와 ReLU에서 파생된 활성화 함수 Leaky ReLU, ELU 이렇게 세 개의 활성화 함수를 이용해서 SeniorCNN 훈련을 수행한 결과 ELU를 사용하여 훈련 시킨 음성파일이 남성, 여성화자 모두 WER이 가장 낮고 PESQ와 MOS가 가장 높게 나왔다. 이것으로 보아 노인 화자의 음성 강화 및 복원을 위한 SeniorCNN 훈련에서는 활성화 함수 ELU를 사용하는 것이 가장 적합하다는 것을 확인할 수 있었다.

## - 목 차 -

제1장 서론 .....	1
제2장 선행 연구 .....	4
2.1 음성 데이터 전처리에 관한 연구 .....	4
2.2 음성 변환에 관한 연구 .....	5
2.3 오디오 스타일변환에 관한 연구 .....	8
2.4 객관적인 음질 평가방법에 관한 연구 .....	10
제3장 SPRocket을 이용한 스타일 음성 생성 .....	13
3.1 실험 데이터 .....	13
3.1.1 고령 화자의 음성 특성 .....	13
3.1.2 서울말 낭독체 발화 말뭉치 .....	14
3.2 음성 변환 수행 .....	15
제4장 SeniorCNN과 활성화 함수 .....	18
4.1 CNN의 정의 및 용도 .....	18
4.2 SeniorCNN .....	19
4.3 활성화 함수 .....	22
4.3.1 ReLU .....	24
4.3.2 Leaky ReLU .....	25
4.3.3 ELU .....	26
제5장 SeniorCNN 훈련 및 후처리 .....	27
5.1 실험의 개요 .....	27
5.2 실험 환경 .....	28
5.3 SeniorCNN의 훈련과정 .....	28
5.4 SeniorCNN의 훈련결과 .....	29
제6장 결과 분석 .....	32
6.1 단어 오류율 측정 결과 .....	32



6.1.1 원시 데이터의 단어 오류율 .....	32
6.1.2 스타일 음성의 단어 오류율 .....	34
6.1.3 활성화 함수별 단어 오류율 .....	34
6.2 PESQ-MOS 측정 결과 .....	36
6.2.1 PESQ-MOS 매핑 함수 .....	36
6.2.2 원시 화자의 음성과 스타일 음성의 PESQ-MOS .....	38
6.2.3 활성화 함수별 PESQ-MOS .....	40
 제7장 결론 .....	 42
 참고문헌 .....	 45
외국어초록 .....	50



## - 그림 목 차 -

<그림 2-1> 음성 변환 기술의 개요 .....	6
<그림 2-2> 음성통계기반 음성 변환 시스템 SPRocket의 작동 과정 ..	7
<그림 2-3> 이미지에서의 스타일변환 .....	9
<그림 2-4> 오디오에서의 스타일변환 .....	10
<그림 2-5> PESQ 척도를 구하는 과정의 블록도 .....	11
<그림 4-1> CNN의 구조 .....	18
<그림 4-2> SeniorCNN .....	19
<그림 4-3> SeniorCNN을 구성하는 계층 .....	20
<그림 4-4> SeniorCNN의 상세 구조 .....	21
<그림 4-5> 퍼셉트론 .....	22
<그림 4-6> 활성화 함수의 종류 .....	23
<그림 4-7> ReLU 함수 .....	24
<그림 4-8> Leaky ReLU 함수 .....	25
<그림 4-9> ELU 함수 .....	26
<그림 5-1> 실험 전 과정의 도식화 .....	27
<그림 5-2> SeniorCNN의 훈련과정 .....	29
<그림 5-3> MZ_09의 SeniorCNN 훈련 후 스펙트로그램의 변화 .....	30
<그림 5-4> FZ_05의 SeniorCNN 훈련 후 스펙트로그램의 변화 .....	30
<그림 5-5> 1 <sup>st</sup> CNN과 2 <sup>nd</sup> CNN의 손실함수 값의 변화량 .....	31
<그림 6-1> PESQ 값으로부터 MOS를 추정할 수 있는 매핑 함수 .....	37
<그림 6-2> PESQ-MOS의 산점도 .....	38
<그림 7-1> SeniorCNN으로 구현한 음악의 스타일변환 .....	43

## - 표 목 차 -

<표 3-1> 60~80대 남녀의 F0 기술통계 결과 .....	14
<표 3-2> 실험에 사용한 데이터의 속성 정보 .....	15
<표 3-3> 음성 변환 실험 설계 .....	15
<표 3-4> 음성 변환 실험 수행 단계 .....	16
<표 3-5> 실험의 3단계 실행 후 화자 별로 얻어진 파라미터들 .....	17
<표 4-1> SeniorCNN의 하이퍼파라미터 설정 .....	21
<표 5-1> 실험 환경 .....	28
<표 6-1> 고령화자 원시 데이터의 단어 오류율 .....	33
<표 6-2> 젊은 화자 원시 데이터의 단어 오류율 .....	33
<표 6-3> 고령화자 스타일 데이터의 단어 오류율 .....	34
<표 6-4> 활성화 함수별 SeniorCNN 훈련결과의 단어 오류율 .....	34
<표 6-5> Leaky ReLU SeniorCNN의 음성인식 결과 .....	35
<표 6-6> MZ_09의 PESQ, MOS 평균값 .....	39
<표 6-7> FZ_05의 PESQ, MOS 평균값 .....	39
<표 6-8> ReLU로 훈련 시킨 음성의 PESQ, MOS 평균값 .....	40
<표 6-9> Leaky ReLU로 훈련 시킨 음성의 PESQ, MOS 평균값 .....	40
<표 6-10> ELU로 훈련 시킨 음성의 PESQ, MOS 평균값 .....	40

## 제1장 서론

인간과 기계의 상호작용은 오랫동안 탐구해 온 연구주제이다. 음성 인터페이스의 진화로 오늘날 전 세계적으로 음성 공학 기술이 우리 생활 깊숙이 자리 잡아 가고 있다. 일례로 아마존의 알렉사, 구글의 시리 등 지능형 음성 기반 비서 서비스 및 챗봇 시스템은 자연스럽고 직관적이다. CUI란 Conversational User Interface의 약자로 컴퓨터 등 기계를 사용하면서 ‘자연스러운 인간과의 대화’를 모방하는 방식이다. 그런데 이러한 CUI는 사전적인 의미의 대화(conversation)와는 구별되는 특징을 가지지만 컴퓨터의 이해 능력을 높이기 위해 자연어 이해(NLU, Natural Language Understanding), 인공지능(AI, Artificial Intelligence), 머신러닝(ML, Machine Learning), 딥러닝(DL, Deep Learning)을 활용하면 인간과 인간의 상호작용을 대체할 수 있다는 것은 매우 먼 미래의 일만은 아니다[1].

이러한 음성 인터페이스는 개념적으로는 매우 간단한 기계-인간 상호 간의 소통을 위한 인터페이스이지만 현재 기술로는 사용자가 키보드나 마우스 등으로 직접 기계에 정보를 입력하는 것보다 오류가 발생할 확률이 높다. 특히 음성은 소음에 민감하므로 오차율이 커서 이러한 환경적인 장애를 극복하고 잡음을 보정 할 수 있는 방법론이 중요한데 하드웨어의 진화를 토대로 최근 비약적으로 발전하는 딥러닝이 음성 인식에서 소음과 음성을 분리하는 작업을 수행해 음성 인식을 증가에 커다란 공을 세우고 있다[2].

음성처리 기술은 매우 복잡하고 그의 적용방식도 난해하여 단순한 물리현상으로 설명할 수 없고 아직 까지는 어떠한 음성 인터페이스에도 통용되는 방법론은 존재하지 않는 것이 현실이지만 상용화에는 큰 기술적 장벽이 있었던 CUI는 점차 우리 생활 어느 곳이나 깊숙이 자리 잡아 사람들 일상생활의 편의를 도모하고 있다.

저출산, 고령화로 인한 여러 가지 사회문제가 대두되는 것은 이미 전 세계적인 추세이다. 한국은 2000년 고령화 사회로 들어선 지 17년만인 2017년에 65세 이상 고령 인구가 14.2%인 711만 5000명에 달해 ‘고령사회’에 진입했다[3].

65세 이상의 고령 화자의 음성이 가지는 공통적인 특성을 분석한 사례[4]를 요약

하면 사람의 음성은 그의 건강 상태를 판단할 수 있는 척도로 사용될 수 있다. 사람은 나이가 들어갈수록 대체로 목소리가 가늘어지고, 그 세기도 작아지며 발화하는데 숨이 차는 것 같은 듣기 편하지 않은 음성으로 변한다.

노인성 음성 질환을 앓는 사람들에게는 젊은 나이에도 이런 증상이 나타날 수 있다고 한다. 연약하고 메마른 음성으로 변화하는 음성으로 인해 심한 경우 우울증, 사회적 고립 등의 문제가 유발되며, 본질적으로 이러한 음성 질환은 코, 귀 등의 이비인후과적인 치료와 함께 필요하면 정신 분석학적 심리치료 등도 진행해야 할 것이다.

두꺼운 사용 설명서를 숙지하지 않고도 제품을 쉽게 사용할 수 있게 하려고 생활 가전제품 등에 이용되는 음성인식은 거동이 불편하고 기계 조작이 자유롭지 못한 독거노인과 같은 고령자와 장애인 등 사회적 약자에게 특히 유용한 기술이다. 그러나 오늘날의 음성인식은 20~30대의 성량이 풍부하고 발음이 좋은 젊은 화자의 데이터를 기반으로 설계되어 있어 진정으로 음성기술이 필요한 사람들은 음성 인터페이스의 편리함을 누리지 못하는 것이 현재 음성 공학 분야에서 극복해야 할 난제이다.

이에 본 연구는 고령 화자의 음성을 적절히 변환함으로써 음성인식 성능을 개선할 수 있도록 하는 방안을 모색하려고 진행하였다.

이와 관련된 선행 연구로 고령 화자의 음성과 젊은 화자의 음성 80개를 비교 분석한 연구가 있다[5]. 참고문헌 [5]의 연구는 고령 화자의 음성은 발화 속도가 느리고, 음절 간 묵음 구간이 길며, 이해하기 어렵다는 특성을 이용하여 묵음 구간을 줄이고, 포만트 주파수 대역대(Formant Frequency Band)를 추가하여 음성 에너지를 증가시키는 전처리를 통해 음성 인식률을 높인 것이다.

또한, 고령 화자의 음성 데이터 특성을 참작하여 CNN을 이용해 고령화자 음성 데이터의 음절 구간을 분류하여 발화율 조정을 통해 음성 인식률을 향상한 연구도 있다[6]. 그러나 이 연구는 모두 고령 화자의 음성만을 가지고 전처리를 시도한 것으로, 다른 음성 데이터와의 융합을 시도해 음성 인식률을 증가시키는 실험을 시도한 것은 아니다. 이는 고령 화자의 음성 이외에 사망자의 음성 복원이나 신호가 약한 소리 데이터의 증강을 위한 응용에는 한계가 있다.

이에 본 논문에서는 고령 화자의 음성이 가지는 약점을 극복해 음성의 품질을 개선함과 동시에 음악 데이터의 전처리 분야에도 적용 가능한 오디오 스타일변환을 통한 고령화자 음성 데이터의 증강 기법에 대해 제안한다.



## 제2장 선행 연구

### 2.1 음성 데이터 전처리에 관한 연구

음성연구에 있어 음성 데이터 자체에 대한 전처리는 음성인식, 음성합성 등의 기술 적용 시 동일한 데이터를 사용하여 더욱 나은 연구결과를 도출하기 위해 선행해야 할 과정이다. 데이터를 머신러닝 알고리즘을 이용해 처리하기 전에 훈련에 적합하도록 처리를 해 주는 것으로 훈련 데이터의 질은 빅데이터 및 딥러닝 연구에 필수적이다. 음성인식 및 음성합성의 연구 진행 시에 음성 데이터 자체에 잡음이 많이 섞여 있거나, 음성이 발화 중간에 끊어지거나, 음성 데이터와 대본이 맞지 않는 경우 본연의 목적에 맞는 결과를 도출할 수 없다. 음성 데이터는 음성 신호가 가지고 있는 정보를 최대한 유지하면서 전체 데이터의 크기를 최소화해야 효율적으로 연산을 수행할 수 있다. 만약에 훈련하고자 하는 데이터의 집합에 결측치(missing value)나 이상치(outlier)가 많이 들어 있다면 딥러닝의 최종 결과에 중대한 영향을 미칠 수 있으므로 전처리는 필수적으로 수행해야 하는 과정이다.

음성 데이터의 전처리는 음성 신호로부터 의미 있는 음성 구간을 검출하고, 음향학적인 파라미터로 변환하는 제반 과정을 의미하며 음성인식에서는 음성과 배경 잡음을 구별하는 음성 구간검출, 음성 신호를 preemphasis 필터를 사용하여 고주파 향의 영향을 높이는 방법 등이 있다[7]. 최근의 전처리 방법은 시간별로 정렬하는 방법(time alignment, 여기에는 선형 시간 정렬과 동적 시간 뒤틀림 기법이 있다) 이외에도 음성 자취 구간검출(trace segmentation) 등의 기법이 있다[8]. 또한, 음성인식과 음성합성 모두 대본에 오·탈자가 없어야 하고 음성 작업에 불필요한 문장부호(., ,, “”, ‘’ 등)와 외래어, 숫자와 특수 기호 등이 발음되는 문자로 명시되어 있어야 딥러닝 알고리즘을 실행하는데 유의미한 결과를 도출해 낼 수 있다. 또한, 음성 데이터 대본의 끝 문자(EOL, End Of Line) 형식이 UNIX 형식인지, Windows 형식인지 아니면

MAC 형식인지에 따라서도 음성 데이터 훈련이 진행되기도, 진행되지 않기도 하므로 해당 데이터의 특성을 상세히 분석해 훈련 목적에 적합하게 설정해야 한다.

2019년 4월 발표된 Google Brain의 연구[9]에 따르면 음성인식을 위한 훈련 데이터를 증강 시키는 방법으로 음성인식 성능의 개선이 가능하다. Google Brain팀의 연구 결과에 따르면 음성 데이터의 특징을 왜곡하고, 주파수 채널을 가리는 방법, 시간 단계를 가리는 방법 등으로 데이터를 증강 시킨 것만으로 다른 언어 모델 사용 없이도 LibriSpeech 데이터베이스의 단어 오류율(WER, Word Error Rate)을 기존의 7.5%에서 6.8%로 낮췄다고 한다. 이를 통해 데이터 자체의 전처리만으로 그 성능이 향상될 수 있다는 것을 알 수 있다.

## 2.2 음성 변환에 관한 연구

음성합성(音聲合成, speech synthesis)은 말소리의 음파를 기계가 자동으로 만들어 내는 기술이다. 이는 모델로 선정된 한 사람의 말소리를 녹음하여 일정한 음성 단위로 분할한 다음에 부호를 붙여 합성기에 입력하였다가 지시에 따라 필요한 음성 단위만을 다시 합쳐 말소리를 인위적으로 만들어 내는 기술이다. 훈련 시킨 데이터가 있다면 사용자는 음성으로 출력하고자 하는 문장 혹은 단어를 입력하여 음성파일로 추출할 수 있다. 최근에는 이와 같은 보통의 음성합성(TTS, Text To Speech) 이외에도 훈련한 음성의 음색을 변환하는 ‘음성 변환’(voice conversion) 기술도 나날이 발전하고 있다. 이는 소위 음성 변환기(voice changer)라고도 할 수 있는데 이러한 음성 변환 기술을 사용하면 자신이 말하는 내용을 유명인의 목소리로 변환할 수 있으므로 게임, 엔터테인먼트 산업이나 개인용 맞춤형 음성 서비스에서 사용할 수 있을 뿐만 아니라 사고나 질병으로 목소리를 잃은 사람들의 음성 복원에도 이 기술이 활용될 가능성을 보이는 잠재력이 매우 큰 기술이다[10].

<그림 2-1>은 음성 변환 기술의 개요를 나타낸다. <그림 2-1>을 통해 음성 변환 기술에 대하여 설명하면 원시 화자인 좌측 화자가 “Welcome to Seoul, Korea! Here is Korea National Open University.”라고 이야기를 해도 목표 화자인 우측 화자의 목소리



로 발화할 수 있도록 하는 것이다. 이러한 경우에 원시 화자와 목표 화자가 모두 “Welcome to Seoul, Korea! Here is Korea National Open University.”라고 발화했다면 병렬 말뭉치(parallel corpus)를 이용한 음성 변환, 원시 화자가 “Welcome to Seoul, Korea! Here is Korea National Open University.”라고 발화했다더라도 목표 화자가 “He sounds like me!”라고 발화했다면 비병렬 말뭉치(non-parallel corpus)를 이용한 음성 변환이다.



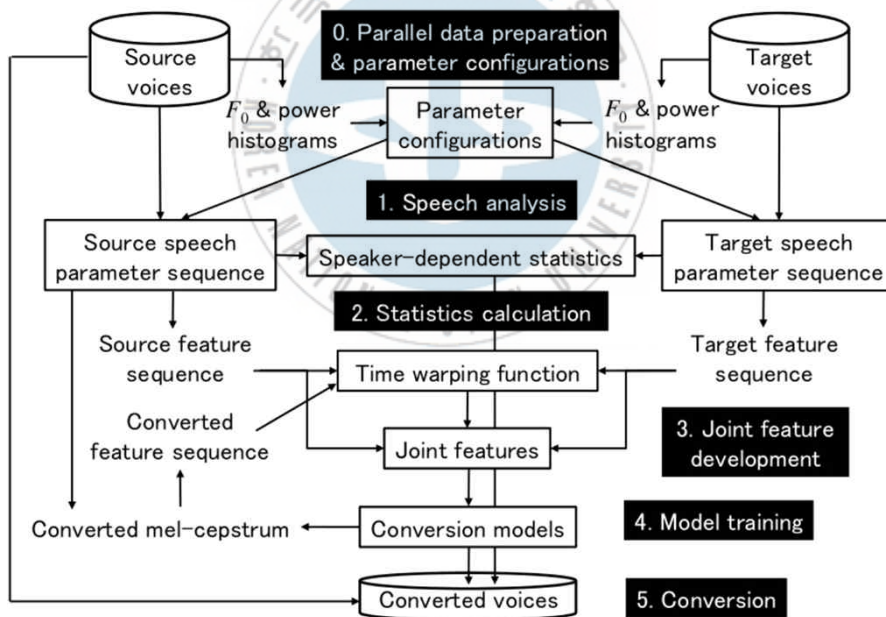
<그림 2-1> 음성 변환 기술의 개요

음성 변환하고자 하는 원시 화자의 발화 내용이 목표 화자의 발화 내용과 항상 같을 수는 없으므로 비병렬 말뭉치를 이용하여 수행하는 음성 변환 기술이 주목을 받고 있다. 비지도 학습 알고리즘과 비병렬 말뭉치를 이용한 음성 변환 연구는 발화 내용이 동일한 병렬 말뭉치를 사용할 때보다 그 활용 분야가 매우 넓어 이의 성능을 개선하기 위한 많은 연구가 진행되고 있다[11-13]. 이러한 노력이 계속되고 있음에도 현재 기술 수준은 i) 음성 변환을 할 때마다 그 결과가 일정하지 않고, ii) GAN(Generative Adversarial Network, 적대적 생성 신경망)과 같은 비지도 학습을 이

용할 경우 훈련 시간이 매우 길어지고 훈련이 제대로 진행되기가 어려우며, iii) 음성 변환하고자 하는 원시 화자와 변환 목표인 목표 화자의 대화 내용이 같아야 음질이 더 좋다는 한계를 갖는다. 현재 기술 수준의 음성 변환에서는 비병렬 말뭉치보다 병렬 말뭉치를 사용하여 실험한 결과가 훈련이 쉽고 추출된 음성의 음질도 전반적으로 좋다는 것을 확인할 수 있다.

SPRocket[14-15]은 병렬 말뭉치를 대상으로 하는 음성 변환 시스템으로 파이썬 (python)으로 구현되었다. 사용자는 시스템의 단계별로 훈련 데이터를 이용하여 기본 주파수  $F_0$ (Fundamental Frequency)의 최솟값, 최댓값을 구하고 경계에 있는 값 (threshold)을 측정하여, 원시 화자 음성의 특성을 자세히 분석하여 최종적으로 목표 화자의 특징을 가지는 음성으로 변환을 수행할 수 있도록 한 시스템이다.

<그림 2-2>는 SPRocket의 전체적인 음성 변환 과정을 보인다.



<그림 2-2> 음성통계기반 음성 변환 시스템 SPRocket의 작동 과정 [16]

SPRocket은 가우시안 혼합모델(GMM, Gaussian Mixture Model) 기반 음성 변환 시

시스템(voice conversion system)이다. 공개된 SPRocket은 음성 변환을 CPU만으로 수행하며 별도의 딥러닝 알고리즘은 적용하지 않았기 때문에 변환된 음성을 얻기 위해 여러 장의 GPU나 대용량의 CPU 등 막중한 하드웨어 장비와 어려운 훈련과정 없이도 음성 변환된 데이터를 출력할 수 있다는 점에서 사용자 친화적이고 실용적인 시스템이다.

프로그램을 실행하면 훈련 데이터로 넣은 음성 데이터 중에서 일부분이 GMM 기반의 음성 변환된 음성파일과 보코더 없이 음성 변환된 파일, 이렇게 두 종류의 결과가 출력된다.

## 2.3 오디오 스타일변환에 관한 연구

스타일변환(style transfer)[17]이란 원래는 음성 데이터가 아닌 이미지에 유명 화가의 화풍을 입혀 새로운 스타일의 이미지를 만들어 내는 것에서 출발한 개념이다. 이미지 데이터가 많지 않아도 짧은 시간에 변환하려고 하는 스타일 인자와 변환의 대상인 원시 데이터의 인자를 조정하여 다양한 디자인을 창출해 낼 수 있는 딥러닝 기법이다. 이는 화풍변환이라고도 하며 이미지에 보이는 대상의 위치·구성은 유지하면서 이미지의 화풍(스타일)만 다르게 변환하는 처리이다.

아래의 <그림 2-3>은 화풍변환의 예를 보여준다. <그림 2-3>의 왼쪽 위의 의자 그림이 스타일변환을 수행할 원시 데이터, 왼쪽 아래의 그림은 저명한 화가 고희의 작품인 ‘별이 빛나는 밤’으로 스타일 데이터이다. 두 개의 이미지 파일을 이용하여 L. A. Gatys 등의 연구[17]에서 착안한 알고리즘으로 스타일변환을 수행한 결과가 우측의 의자 그림이다. <그림 2-3>에서 알 수 있듯이 이미지에서의 화풍 스타일변환의 결과물은 원본인 의자 그림에 고희 작품의 색채와 질감의 특성이 반영된 것임을 확인할 수 있다.

L. A. Gatys 등의 연구[17]를 시작으로 화풍변환은 빠르게 발전하였으며 오늘날 웹이나 모바일 앱에서도 매우 활발하게 이용되는 기술이다. 이미지는 오디오와는 달리 원 이미지(content image) 위에 스타일 이미지의 분위기만 나타나면 되고, 스타일 이

미지가 넓게 퍼져도 원래 변환하고자 했던 이미지와 조화를 이룬다면 아름다운 한 폭의 작품이 될 수 있다.



<그림 2-3> 이미지에서의 스타일변환

하지만 본래 음성 데이터와 구분되는 잡음(noise)은 데이터에서 없애야 하는 부분으로 음악이나 음성과 같은 오디오 데이터를 이용한 스타일변환은 쉽지 않은 작업이다. 음성 데이터는 시계열 데이터로 연속적이기 때문에 스타일변환으로 데이터 손실이 발생할 수밖에 없는 구조이다. 이미지나 동영상에서 데이터의 손실이 아름다운 또 하나의 작품으로도 될 수 있는 것과는 다른 음성 데이터의 특징으로 인해 스타일변환으로 발생하는 많은 잡음은 음성 데이터를 치명적으로 오염시켜 사용할 수 없도록 하는 요인이 되기도 하므로 이를 오디오에 적용하는 연구는 지금까지 많은 시도가 있었으나 그 결과물은 새로운 음성 혹은 음악 데이터로 쓸 수 없는 수준에 그치는 것이 많았다.

Eric Grinstein 등의 연구[18]에 따르면 연구결과인 유명 가수 Eminem의 노래를 이용한 스타일변환, 인기 가요에 링컨의 게티즈버그 연설을 스타일로 입힌 오디오 스타일변환 실험은 음성과 음악 모두에서 뚜렷한 결과를 나타내지 못하였다. 또한, 교향악단의 연주에 오페라 가수의 노래를 스타일로 입힌 Alish Dipani의 연구[19] 역시

최종 결과는 가수의 노래인지 교향악단의 합주인지 알 수 없는 형태의 오디오가 생성되는 것에 그쳤다.

<그림 2-4>는 오디오에서의 스타일변환이 갖는 특징을 직관적으로 나타내었다. 좌측의 원시 화자의 발화 내용은 스타일변환 하고자 하는 음성 데이터, 우측의 목표 화자의 발화 내용은 변환하고자 하는 대상이 되는 음성 데이터가 된다. 즉, 오디오에서의 스타일변환이란 원시 화자의 음성에 목표 화자의 음성이 부가되는 결과물을 산출하는 것을 말한다.



<그림 2-4> 오디오에서의 스타일변환

## 2.4 객관적인 음질 평가방법에 관한 연구

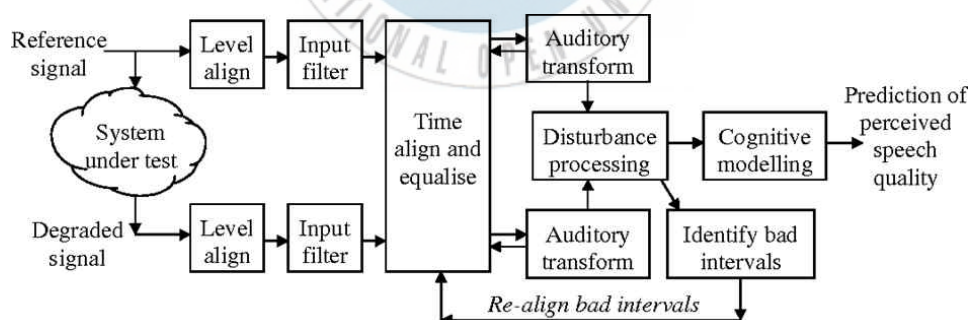
본 논문에서는 CNN과 음성 변환을 이용하여 고령 화자의 음성 전처리에 사용할 것이므로 변환된 음성이 변환 전 음성 대비 음질의 변화가 어떤지를 판단하는 객관적인 척도가 필요하다.

기존의 객관적인 음질 측정 방법으로는 시간 영역에서의 척도로 신호 대 잡음 비

(SNR, Signal to Noise Ratio)가, 주파수 영역에서의 척도로 두 음성 사이의 스펙트럼 거리, 캡스트럼 거리 등이 사용되었다. 그러나 이러한 단순한 시간 영역 또는 주파수 영역에서의 거리로는 객관적 음질 평가를 위한 만족스러운 결과를 얻을 수 없고, 바크 스케일(Bark scale)과 같은 사람이 음성을 지각하는 영역으로의 변환을 통해서 거리를 구하는 척도가 주로 이용되었다.

이보다 객관적인 음질 평가를 위해서는 원래의 음성을 기준으로 하여 평가하고자 하는 음성과의 차이를 나타내는 척도를 이용하는 것이 바람직하다. 이에 본 논문에서는 통신에서의 통화 품질을 측정하기 위한 용도로 개발된 PESQ(Perceptual Evaluation of Speech Quality)와 이를 주관적인 척도인 MOS(Mean Opinion Score)로 변환한 계산식을 적용하여 아무 처리를 하지 않은 본래의 고품질 화자의 음성을 참조 데이터로 하여 PESQ와 MOS값을 SPRocket의 결과 음성과 CNN 훈련으로 생성된 음성 모두에 대해 구해서 그 값을 비교하였다.

PESQ는 전화 대역의 음성 신호를 대상으로 개발된 음성품질 측정 알고리즘이다 [20]. 이는 음성 코덱, 변동적인 지연, 필터링, 패킷이나 셀 손실 및 채널손실을 가지는 시스템에 적용할 수 있도록 고안된 표준이다. <그림 2-5>는 PESQ 척도를 구하는 과정의 블록도를 나타낸다.



<그림 2-5> PESQ 척도를 구하는 과정의 블록도 [21]

PESQ를 구하는 과정은 다음과 같다.



1. 입력 음성 신호와 시스템을 통과한 출력 음성 신호를 주관적 음질 평가에서 일반적으로 사용하는 레벨과 비슷하게 하려고 표준청각 레벨을 기준으로 입력신호의 레벨을 정렬하는 과정과 수화기의 대역 통과 특성을 고려하기 위한 필터링 과정
2. 왜곡된 음성 신호의 지연을 고려하여 두 음성 신호의 시작점을 찾아 시간 정렬을 수행하고 비 가청음 영역의 신호를 제거하는 과정
3. 원래 음성 신호와 왜곡된 음성 신호의 소리의 강도(loudness density) 차이를 계산하여 소리의 강도가 낮은 부분에 대해서 다시 시간 정렬 과정을 실행하여 더욱 정확한 시간 정렬 과정을 수행
4. 사람의 청각특성을 고려한 지각영역으로의 변환과 인지 과정 모델을 거침으로써 최종적으로 -0.5에서 4.5의 범위를 가지는 PESQ 득점을 산출

임창송 등이 2010년 수행한 연구[21]에서는 한국어 합성음의 객관적인 음성 데이터 음질의 평가방법으로 사용되는 PESQ 척도의 타당성을 알아보기 위한 목적으로 HMM(Hidden Markov Model) 기반 음성합성으로 합성한 합성음에 대한 주관적 평가 실험을 수행하였고, 그 결과인 MOS 및 DMOS(Difference MOS)와 객관적 평가방법인 PESQ를 이용한 득점과의 상관도를 분석하였다. 그 결과 한국어 합성음에 대해 PESQ를 이용한 객관적인 음질 평가 결과는 주관적인 음질 평가 결과인 MOS와는 0.87의 상관도를 보였으며, DMOS와는 0.92 정도의 높은 상관도를 보여 한국어 합성음의 객관적인 음질의 평가방법으로 PESQ 척도가 충분히 사용될 수 있다고 하였다.

본 연구는 전래동화인 ‘선녀와 나무꾼’을 대본으로 하는 한국어 음성을 SPRocket을 사용한 음성 변환, 이렇게 음성 변환된 데이터를 스타일 음성으로 하여 CNN 훈련으로 오디오 스타일변환을 수행하였으므로 선행 연구에 따라 원시 화자 음성 대비 변환한 음성을 PESQ-MOS를 척도로 하여 품질을 평가해 보겠다.

## 제3장 SPRocket을 이용한 스타일 음성 생성

### 3.1 실험 데이터

#### 3.1.1 고령 화자의 음성 특성

고령 화자의 음성은 젊은 연령대의 화자보다 잡음이 많고 희미하며 발음도 부정확한 경우가 많다. 이는 해부학 · 생리학적으로 사람이 나이가 들어감에 따라 혀의 움직임의 범위가 줄어들고 성대 모양도 변형을 일으켜 구강과 성대 부근이 건조해지고 둔해지며 청각장애가 수반되어 오는 영향 등 여러 가지 요인이 있다.

일반적으로 사람은 30세 이후부터 생리적, 심리적, 사회적 측면으로 매년 1%씩 신체의 구조와 기능이 저하되는데 이로 인해 모든 신체는 기능적 결손을 차차 초래하게 된다[22]. 노화가 진행될수록 사람의 신체 기관 중 조음 기관인 후두 및 구강, 호흡 관련된 기관은 전체적으로 특별한 병력이 없어도 노화만으로 생리적인 변화가 발생한다. 그 예로는 성대 휨(bowing), 성대위축(atrophy), 불완전 성문폐쇄(glottal incompetence) 등이 있다. 이러한 신체적 노화 현상으로 나타나는 대표적인 증상이 고령 화자에게서 볼 수 있는 일반적인 특징으로 음도 및 강도의 제한, 약한 음성(weak voice), 기식음(breathiness), 원목소리(hoarseness) 등이다[23].

노화로 인해 음성에 이러한 변화가 오면 일상생활에서의 의사소통 능력도 떨어질 뿐만 아니라 현재 음성을 인터페이스로 하는 기술들이 거동이 불편하고 기계에 익숙하지 않은 노인들에게 필요함에도 이에 접근하지 못하는 문제가 발생해 노인들의 삶의 질을 떨어뜨리는 주요한 원인이 된다.

특별한 이비인후과적인 질환이 없음에도 한국의 정상 노인 음성의 기본 주파수를 고찰한 연구[24]에 따르면 서울과 경기 지역에 거주하는 60~89세까지 건강한 노인 남성 207명, 여성 199명 총 406명을 대상으로 한 연구에서 산출한 노인 남녀의  $F_0$ (기본 주파수, fundamental frequency) 기술통계 결과는 <표 2-1>과 같다.



<표 3-1> 60~80대 남녀의 F0 기술통계 결과

	남성		여성	
	평균	표준편차	평균	표준편차
60대	139.93	9.71	189.91	11.24
70대	143.65	13.56	181.39	14.69
80대	148.25	16.59	185.17	17.99
총계	143.95	13.94	185.42	15.29

김선해 등이 진행한 연구[24]의 결론에 따르면 고령의 남성화자 음성의 기본 주파수  $F_0$ 는 60~80대에 걸쳐 서서히 상승하는 반면 여성은 60대에서 70대까지는  $F_0$  값이 하강하다가 80대가 되면 다시 상승하는 경향을 보인다.

### 3.1.2 서울말 낭독체 발화 말뭉치

본 연구에서는 국립국어원에서 2004년도에 제작하고 2005년도에 민간에 공개한 ‘서울말 낭독체 발화 말뭉치’를 실험 데이터로 사용하였다. 이 데이터는 한국소설, 전래동화 등 한국어 특유의 감성이 잘 묻어나는 내용으로 구성된 다 화자 병렬 말뭉치(multi-speaker parallel corpus)이다. 이 데이터를 실험 데이터로 채택한 이유는 다음과 같다. 이 데이터는 한국어 음성 코퍼스로 남성, 여성 화자를 성별과 연령 별로 분류하여 모두 동일한 대본을 낭독하게 한 병렬 말뭉치(parallel corpus), 즉 20대부터 70대까지 연령 별로 다수의 화자가 동일한 대본을 낭독하는 형태로 구성된 데이터이다. 이에 SPRocket의 시스템을 사용할 수 있는 데이터로서의 요건을 갖추었다.

또한, 음성연구가 가능하도록 공개된 한국어 음성 말뭉치가 많지 않은 현실에서 이 데이터는 화자의 성별, 나이대별로 문장 단위로 녹음되어 별도의 편집이 필요 없고 고연령층 화자와 젊은 연령층 화자의 음성을 모두 사용할 수 있으며 메타데이터의 속성 정보가 자세히 기재되어 있어 실험에 사용하기 유용하여 이 데이터를 선택하게 되었다. 다만, 스튜디오 환경에서 녹음한 것이라곤 하지만 고령 화자의 경우, 그들의 음성이 가지고 있는 특성에 따라 잘 들리지 않는 단점은 있었다.

실험에 사용한 각 화자의 ID와 화자별 속성 정보는 <표 3-2>에 정리하였다.

<표 3-2> 실험에 사용한 데이터의 속성 정보

화자 기호	녹음 연도	성 별	녹음 당시 나이	출생지	성장지	부친 출신지	모친 출신지
FV01	2003	여	만 23세	서울	서울	서울	서울
FV13	2003	여	만 29세	경기 안양	경기 안양	경기 안양	충남 서산
FZ05	2003	여	만 68세	서울	서울	서울	서울
FZ06	2003	여	만 65세	서울	서울	서울	서울
MV13	2003	남	만 25세	경기 안양	서울	경남 합천	경기 용인
MW01	2003	남	만 30세	인천	인천	인천	인천
MZ05	2003	남	만 68세	서울	서울	서울	서울
MZ09	2003	남	만 71세	서울	서울	서울	서울

### 3.2 음성 변환 수행

데이터 전처리를 위한 실험 설계는 <표 3-3>과 같이 하여 실험을 진행하였다.

<표 3-3> 음성 변환 실험 설계

Source Speaker			Target Speaker		
Sex	Speaker ID	Age	Sex	Speaker ID	Age
M	MZ05	68	M	MV13	25
				MW01	30
	MZ09	71		MV13	25
				MW01	30
		F	FV01	23	
F	FZ06	65	F	FV01	23
				FV13	29
	FZ05	68		FV01	23
				FV13	29
		M	MW01	30	

남성 화자와 여성 화자를 각각 원시 화자, 목표 화자로 두어 실험을 진행한 결과

원시 화자가 남성이라면 목표 화자도 남성으로, 마찬가지로 원시 화자가 여성이라면 목표 화자도 여성이어야, 즉 동성이어야 음성 변환이 수월하게 진행된다는 것을 확인할 수 있었다.

이러한 특징은 여성 화자를 원시 화자, 남성 화자를 목표 화자로 두어 음성 변환을 수행했을 때에도 마찬가지로 나타났다. 이는 남성과 여성의 음성이 본질적으로 기본 주파수 범위가 다르기 때문으로 추론할 수 있다. <표 3-1>에서 알 수 있듯이 여성 화자의 평균  $F_0$ 가 남성 화자의  $F_0$ 에 비해 약 40 정도 높기 때문이다.

SPRocket System을 사용하기 위해서는 <표 3-4>의 다섯 단계의 절차를 거쳐야 한다.

<표 3-4> 음성 변환 실험 수행 단계

단계	수행해야 할 작업
1	리스트 파일 생성: 음성 데이터를 훈련 데이터와 테스트 데이터로 리스트를 작성하여 나눈다.
2	생성된 파일 조정: 원시 화자와 목표 화자, 그리고 음성 데이터의 표본 데이터를 지정한다.
3	수동 설정: 각 모듈에서 추출되는 $F_0$ 히스토그램과 정규화된 파워 스펙트럼을 분석하여 $F_0$ 의 최댓값, 최솟값 및 경계값을 음성파일별 가지고 있는 yaml 파일에서 수정해야 한다.
4	훈련단계: 음성 통계학적인 GMM 방법으로 데이터를 훈련한다.
5	분석단계: 훈련한 통계 데이터를 기반으로 음성 변환을 수행하고 음성 변환 실행 결과를 분석한다.

원시 화자로 선정한 데이터는 다음과 같다. 남성은 71세의 MZ09가, 여성은 68세의 FZ05가 다른 고령 화자의 음성에 비해서 울리는 소리가 더 많이 나고 낭독한 내용의 전달이 잘되지 않는다고 판단하여 최종적으로 실험은 남성 고령 화자는 MZ09를, 여성 고령 화자는 FZ05를 원시 화자의 음성으로 선정하여 이들의 음성 데이터를 가지고 SPRocket을 사용한 음성 변환 실험을 진행하였다.

실험을 수행한 후 얻어진 파라미터는 <표 3-5>에 정리하였다.

<표 3-5> 실험의 3단계 실행 후 화자 별로 얻어진 파라미터들

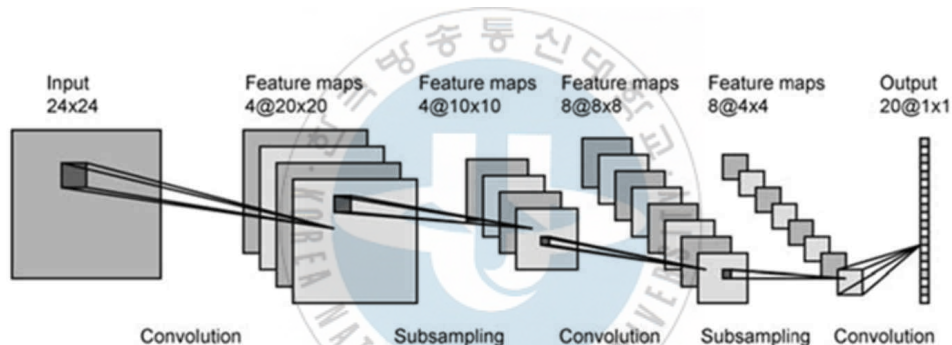
Speaker	Minimum F0 (Hz)	Maximum F0 (Hz)	Power (dB)
MV13	80	190	-25
MW01	80	190	-20
MZ05	45	190	-20
MZ09	45	190	-25
FV01	140	340	-30
FV13	120	340	-15
FZ05	90	290	-25
FZ06	90	240	-20



## 제4장 SeniorCNN과 활성화 함수

### 4.1 CNN의 정의 및 용도

CNN(Convolutional Neural Network)은 Yan LeCun 교수가 1998년 소개한 이래로 널리 사용되고 있는 신경망이다[25]. 특히 이 신경망은 이미지 분야에서 강력한 성능을 발휘하며 최근에는 자연어처리나 음성인식 등에도 활용되며 뛰어난 성능을 보여주는 신경망이다.



<그림 4-1> CNN의 구조 [26]

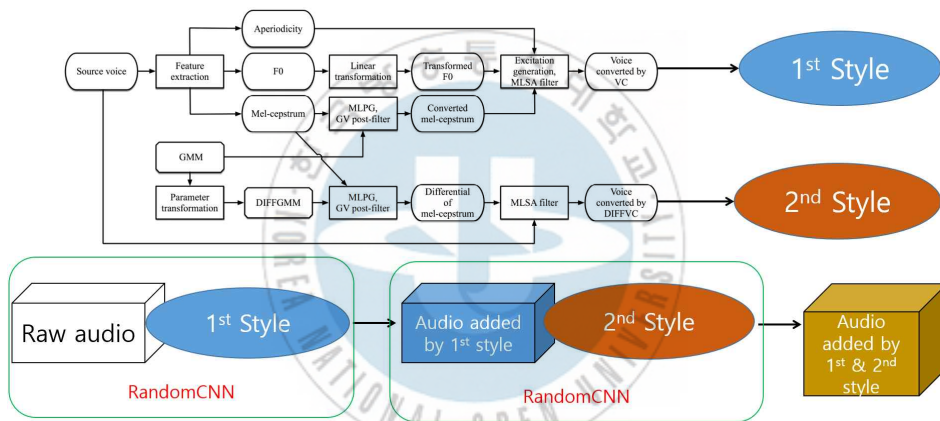
<그림 4-1>의 CNN 구조에서 살펴볼 수 있는 것과 같이 CNN의 가장 큰 특징은 반복되는 합성곱 층(convolution layer)과 풀링 층(pooling layer)이다.

합성곱 층은 이미지에 필터를 적용해 이미지의 특징량을 추출하는 역할을 담당하는 층으로 파라미터의 수는 이미지 크기가 아닌 필터 크기에 의존한다. 그리고 풀링 층은 작은 변화에 민감하지 않도록 하는 역할을 담당하며 이미지를 축소하는 역할을 담당한다. 가장 자주 사용되는 최대 풀링은 입력 데이터를 작은 영역으로 분할할 뿐만 아니라 각 영역의 최댓값을 추출하면서 데이터를 축소한다. 궁극적으로 CNN은 입력 데이터의 성질을 이용해서 파라미터의 수를 줄이는 특징을 가진다고 할 수 있

다.

## 4.2 SeniorCNN

이미지의 화풍변환도 CNN을 기반으로 시작되었고 음성 변환 및 오디오 스타일변환을 적용하는 본 실험의 목표가 원시 화자의 성질을 최대한 유지하면서 스타일 화자의 특징을 원시 화자의 음성에 적용하는 것이므로 오디오 스타일변환을 위한 딥러닝 모델은 CNN을 기반으로 한다.



<그림 4-2> SeniorCNN

SeniorCNN은 오디오 스타일변환에 있어서 CNN 2개 층으로 구현한 GitHub의 공개된 코드[27]를 기반으로 재설계한 것이다. <그림 4-2>는 본 논문연구를 통해 설계한 SeniorCNN의 전체적인 구조를 나타낸 것이다. 이 연구는 고령 화자의 음성 데이터 강화를 통해 고령 화자들이 음성 인터페이스를 유연하게 사용할 수 있도록 하는 것이 목적이므로 본 연구에서 고안한 딥러닝 신경망을 고령자를 뜻하는 영어 단어인 Senior를 붙여 SeniorCNN으로 명명하였다.

<그림 4-2>의 상단에 있는 순서도는 SPRocket의 실행과정을 그리고 있다. SPRocket

을 실행하면 최종적으로 전통적인 GMM 기반의 음성 통계학적인 기법으로 추출한 VC(Voice Conversion) 음성파일과 보코더 없이 음성 변환한 DIFF\_VC 이렇게 두 가지 파일이 나온다. SeniorCNN은 두 단계의 CNN 훈련을 거치도록 설계하였다. 이는 여러 번의 CNN 훈련이 고령화자 음성의 음질 개선에 도움이 되는지를 판단하기 위함이다.

첫 번째 스타일 음성은 통상적인 GMM 기반의 알고리즘으로 음성 변환된 음성을 사용하였고 두 번째 스타일 음성으로는 보코더 없이 음성 변환을 수행하여 얻은 결과물인 DIFF\_VC 음성을 사용하였다.

```
RandomCNN (
  (conv1): Conv2d(1, 32, kernel_size=(3, 1), stride=(1, 1), weights=((32, 1, 3, 1), (32,)) parameters=128
  (LeakyReLU): LeakyReLU(negative_slope=0.2, weights=(), parameters=0
)
```

<그림 4-3> SeniorCNN을 구성하는 계층

<그림 4-3>은 SeniorCNN을 구성하는 계층을 보여준다. <그림 4-3>은 활성화 함수를 Leaky ReLU로 했을 때의 전체적인 구조를 나타내며 ELU, ReLU의 경우 <그림 4-3>의 ‘LeakyReLU’ 부분이 각각 ELU, ‘ReLU’로 바뀐다. LeakyReLU는 dying ReLU 현상을 방지하는 효과가 있다고 알려져 있으므로 SeniorCNN 성능을 최적화하기 위해 유용한 지 확인하기 위해 채택하였다[28].

SeniorCNN은 2D-RandomCNN이다. 2D-RandomCNN의 특징은 <그림 4-3>에서 확인할 수 있다. 빨간색 상자로 표시한 가중치 값이 2차원이고 두 번째 차원의 가중치 값은 난수 발생을 이용하여 임의로 받아 생성하기 때문에 (32, )로 두 번째 요소가 공란이다. 이러한 특성을 갖기 때문에 Random이라는 단어를 사용하여 SeniorCNN의 특징을 표현하였다.

SeniorCNN은 2D-RandomCNN의 훈련을 하이퍼파라미터를 <표 4-1>과 같이 달리하여 두 단계로 훈련 시킨 것이다.

<그림 4-4>는 입력 차원(input shape)과 출력 차원(output shape)을 포함한 SeniorCNN의 상세 구조를 도식화한 것이다. 본 모델은 딥러닝 라이브러리 중 pytorch



를 이용하여 구현하였다. 128개의 모든 파라미터는 훈련되지 않는(non-trainable) 파라미터이다. 즉, 파라미터는 은닉 계층(hidden layer)의 값이다.

<표 4-1> SeniorCNN의 하이퍼파라미터 설정

Hyperparameters	1 <sup>st</sup> CNN	2 <sup>nd</sup> CNN
learning rate	0.002	0.002
style_param	2	1
content_param	1e4	1e2
num_epochs	10,000	20,000
print_every	1,000	1,000
plot_every	1,000	1,000

Layer (type)	Input Shape	Output Shape	Param #
Conv2d-1	[-1, 1, 257, 613]	[-1, 32, 255, 613]	128
LeakyReLU-2	[-1, 32, 255, 613]	[-1, 32, 255, 613]	0
Total params: 128			
Trainable params: 0			
Non-trainable params: 128			
Input size (MB): 0.60			
Forward/backward pass size (MB): 76.33			
Params size (MB): 0.00			
Estimated Total Size (MB): 76.93			

<그림 4-4> SeniorCNN의 상세 구조

SeniorCNN은 원시 화자의 데이터에 목표 화자의 특성을 입히는 오디오 스타일 변환을 위한 CNN이므로 음성인식, 음성합성 등에 사용되는 통상적인 딥러닝 기법과는 다르게 접근하였다. 이는 처음에 설정한 하이퍼파라미터와 활성화 함수로 SeniorCNN의 결과 음성의 품질이 정해진다는 것을 의미한다.

<표 4-1>의 파라미터들을 보면 원시 화자의 특성을 보존하는 역할을 하는

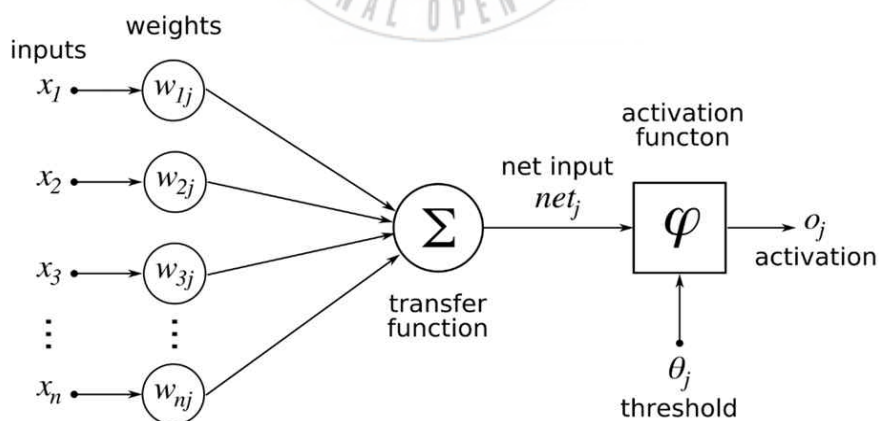


content\_param을 첫 번째 단계에서 훨씬 더 작게 설정했음을 확인할 수 있다. 목표 화자의 특징을 나타내는 style\_param은 두 번째 단계의 CNN 대비 첫 번째 단계의 CNN에서 두 배로 설정하였다. <표 4-1>과 같이 파라미터를 설정한 이유는 오디오 스타일변환을 통해 얻은 결과는 잡음이 많이 발생한다는 선행 연구결과를 고려하여 원시 화자의 음성과 스타일로 사용한 음성을 적당히 융화시키기 위함이다. 원시 화자의 음성이 스타일 음성에 저항하는 데이터로 남아서는 안 되기 때문에 스타일 음성의 성질을 최대한 결과 음성에 반영하여 결과 음성의 품질을 높이기 위해 하이퍼 파라미터를 <표 4-1>과 같이 설정하였다.

스타일 음성으로 SPRocket의 변환 결과를 사용한 이유도 현행 오디오 스타일변환 기술로는 원 음성의 발화 내용을 알아들을 수 있을 만큼의 스타일변환이 이루어지지 않는다는 한계점을 고려했기 때문이다.

### 4.3 활성화 함수

활성화 함수(activation function)란 임계치의 초과 여부를 판단하는 함수로 임계치를 초과하면 1, 아니면 0 (혹은 -1)과 같이 특정 값을 출력하도록 하는 함수이다[29].

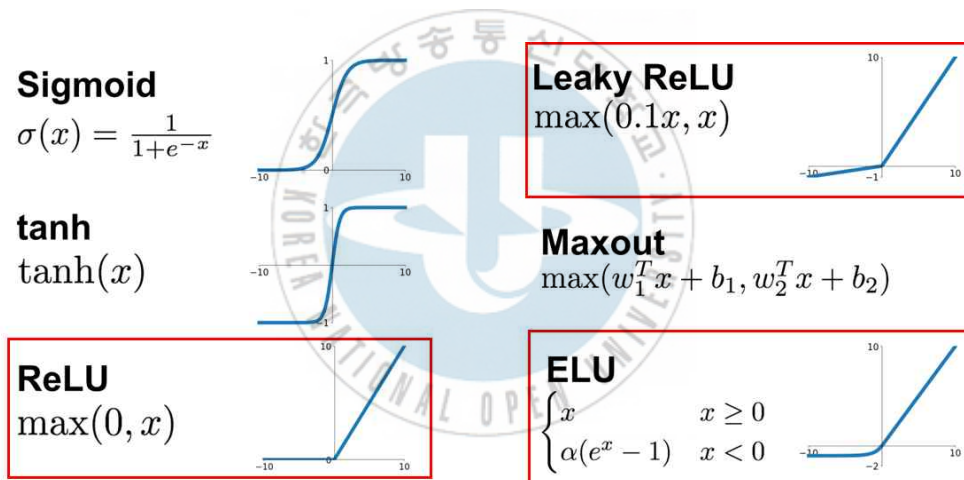


<그림 4-5> 퍼셉트론

이때 임계치를 넘었다면 ‘활성화’ 되었다고 정의한다. 즉, 활성화 함수는 입력한 신호의 총합을 출력신호로 변환하는 함수이자 신경망의 출력을 결정하는 식으로 입력한 신호의 총합이 활성화를 일으키는지를 정하는 역할을 한다.

활성화 함수는 1950년대에 Rosenblatt가 신경세포를 모방하여 개발한 알고리즘인 퍼셉트론(perceptron)에서 그 용도를 명확히 확인할 수 있다. 퍼셉트론은 여러 개의 입력신호를 받아서 하나의 신호를 출력하는 알고리즘으로 <그림 4-5>는 활성화 함수의 역할을 직관적으로 확인할 수 있는 퍼셉트론을 나타낸 것이다.

활성화 함수는 시그모이드(sigmoid), tanh, ReLU 등 굉장히 다양한 종류가 있으나 임계치 초과 여부를 판단하여 어떤 값을 출력할 것인가를 결정한다는 점에서 그의 본질적인 역할은 모두 같다.



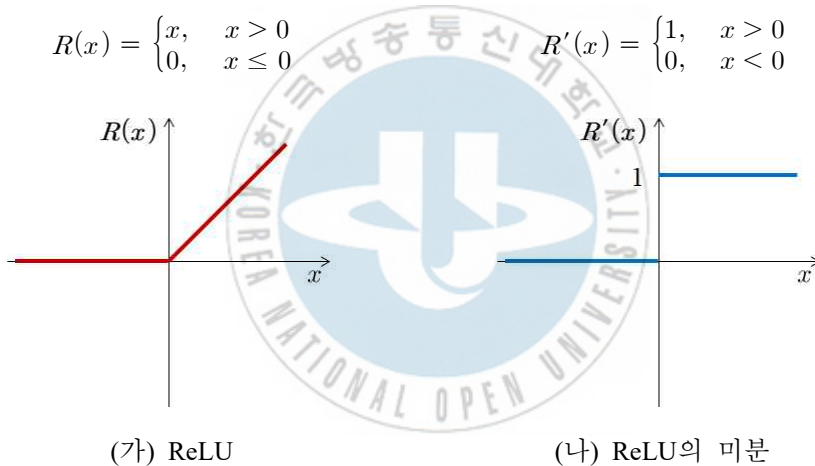
<그림 4-6> 활성화 함수의 종류 [30]

활성화 함수의 종류는 <그림 4-6>과 같다. 본 논문에서는 활성화 함수별로 훈련 후 결과 음성의 음질을 비교하기 위해 <그림 4-6>에서 빨간색 상자로 표시한 ReLU, Leaky ReLU, ELU를 실험에 적용할 활성화 함수로 채택하였다. 채택한 세 가지 활성화 함수는 가장 많이 사용되며 모두 ReLU라는 활성화 함수의 변형이라는 공통점이 있다.

### 4.3.1 ReLU

활성화 함수 ReLU(Rectified Linear Unit)는  $x$ 가 양수이기만 하면 기울기가 1로 일정해 시그모이드나 tanh가 가지는 기울기 값이 사라지는 문제(vanishing gradient problem)를 해결한 함수로써 매우 큰 의의가 있다. 대부분은 시그모이드나 tanh 함수 대비 학습 수렴 속도가 6배 정도 빠르다고 알려져 있으며 미분을 하기에 편리하여 계산 복잡성이 낮다.

<그림 4-7>은 ReLU 함수가 출력값으로 어떤 값을 가지는지 그래프와 함께 잘 보여주고 있다.



<그림 4-7> ReLU 함수

ReLU 함수는 시그모이드 함수가  $x$ 가 음수일 때 0이 아닌 값을 가지므로 밀도 있게 값이 존재하게 되어 생기는 문제(dense representation)를 해결한 함수로 값을 희소하게 표현하는(sparsity representation) 성질을 보이는 강점이 있다.

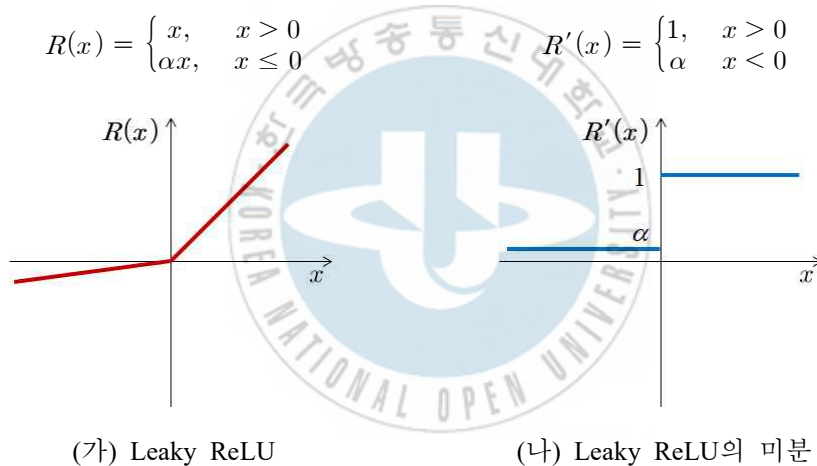
이렇게 기존 활성화 함수의 한계점을 극복한 ReLU 함수도 1) 신경망 모델의 은닉 계층(hidden layer)으로만 사용되는 한계점, 2) 입력값이 음수이면 무조건 0이 되므로 훈련을 지속해도 가중치가 업데이트되지 않는 문제점 (dying ReLU problem)으로 죽

은 뉴런(dead neuron)의 원인이 될 수 있다는 점, 3) 활성화가 폭발하는 문제(blow up activation)의 위험을 안고 있다는 단점이 있다.

3)의 한계점은 ReLU 값의 범위 때문에 생기는 문제이다. ReLU 값이  $[0, \infty)$  라는 의미는 그 값이 무한대로 증가할 수 있으므로 활성화 함수의 기능을 제대로 하지 못할 수도 있다.

### 4.3.2 Leaky ReLU

활성화 함수 ReLU의 변형인 Leaky ReLU는 <그림 4-8>과 같이 정의한다.



<그림 4-8> Leaky ReLU 함수

이 함수는 활성화 함수로써 성능이 가장 뛰어나 널리 쓰이는 ReLU 함수가 입력값이 음수일 때 훈련을 지속해도 데이터가 갱신되지 않는 dying ReLU 현상을 방지하기 위해 기울기를 0이 아닌 0.01과 같이 미세한 값을 주어서 이 문제점을 극복하기 위해 시도한 것이다.

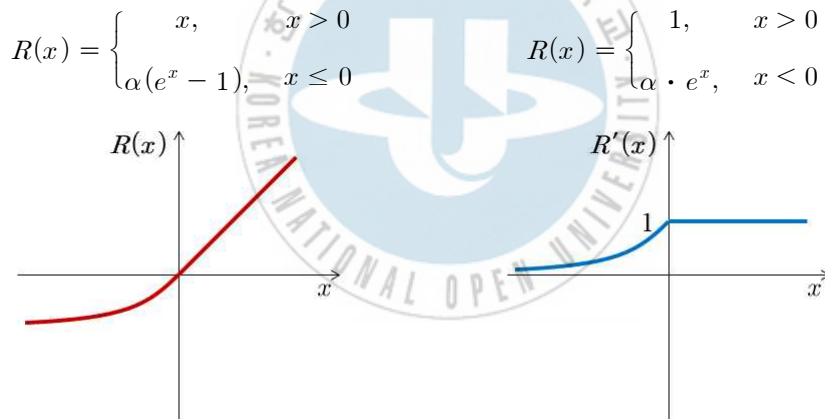
그러나 이러한 Leaky ReLU는 선형성을 가지므로 복잡한 분류 모델 등에는 사용될 수 없고, 오히려 tanh나 시그모이드 함수보다도 성능이 저하되는 경우도 발생한다는

문제점이 있다.

### 4.3.3 ELU

ELU(Exponential Linear Unit)는 ReLU가 가파르게 0으로 수렴하는 것과 달리 입력값이 음수일 때도 기울기가 서서히 변하므로 ReLU가 가지는 단점을 보완하는데 사용되는 활성화 함수이다. <그림 4-9>의  $x \leq 0$ 인 구간에서 이러한 성질을 확인할 수 있다. 입력값이 음수일 때 입력값의 절대값이 커질수록  $e^x$ 은 0에 수렴하므로 출력값은  $-\alpha$ 에 가까워진다.

Leaky ReLU 함수와 마찬가지로 ELU 함수는 기울기가 음수일 때도 출력값을 가진다. <그림 4-9>는 ELU 함수의 성질을 나타낸다.



(가) ELU

(나) ELU의 미분( $\alpha = 1$ )

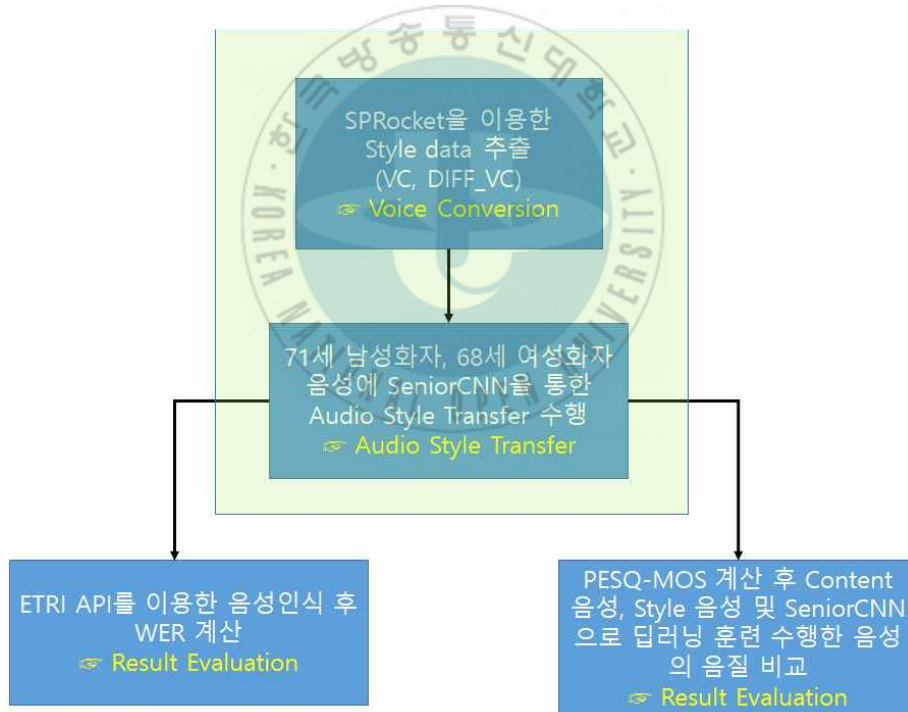
<그림 4-9> ELU 함수

ELU는 ReLU와 마찬가지로 출력값의 범위가  $[-\alpha, \infty]$  이기 때문에 ReLU가 가지는 활성화가 폭발하는 문제점을 가지는 단점이 있다.

## 제5장 SeniorCNN 훈련 및 후처리

### 5.1 실험의 개요

본 논문의 실험은 <그림 5-1>의 과정을 거쳐 음성 변환 결과물을 얻고 그 결과물을 이용하여 오디오 스타일변환 기법을 적용하도록 설계하였다. 투명한 연두색 상자 안의 전 과정이 음성인식 전에 음성 데이터를 증강하기 위한 전처리 과정에 해당한다.



<그림 5-1> 실험 전 과정의 도식화

음성 변환과 스타일변환 모두 결과 자체로서 의미가 있는 딥러닝을 이용한 음성연구 분야로 그 수행 결과를 힘이 없고 울리는 소리가 많이 나는 고령화자 음성의 전처리에 사용한 선행 연구는 조사 결과 발견하지 못했다.

## 5.2 실험 환경

본 논문의 실험은 개인용 노트북으로 진행하였으며 실험을 수행한 전체적인 하드웨어의 사양은 다음과 같다.

<표 5-1> 실험 환경

종류	사양
CPU	Intel Core i9 8950HK
GPU	RTX 2070 Max-Q
NVIDIA CUDA / CUDNN	CUDA 10 / CUDNN 7.5
SSD	1TB
Operating System	Ubuntu 16.04

## 5.3 SeniorCNN의 훈련과정

<그림 5-2>는 SeniorCNN의 훈련과정을 나타낸다. 각각 2개의 CNN 훈련을 총 4번 실행하도록 설계하였으므로 파이썬 실행 창에서는 <그림 5-2>와 같이 훈련이 100% 되는 로그가 두 번 출력됨으로써 한 음성 데이터의 오디오 스타일변환 과정을 종료하게 된다.

실험은 ReLU, Leaky ReLU, ELU 이렇게 훈련 코드의 활성화 함수를 바꿔가며 세 가지로 진행하였다.

최종 실험결과는 각 활성화 함수별로 MZ\_09 음성 66개, FZ\_05 음성 66개가 나왔다. 이는 첫 번째 CNN 훈련과 두 번째 CNN 훈련의 결과 음성이 각각 33개로 두 결과 음성을 합친 개수이다.



```

torch.Size([1, 1, 257, 810])
torch.Size([1, 1, 257, 810])
1000 10.0% 2m 35s content_loss:1.705243 style_loss:0.819888 total_loss:2.525131
2000 20.0% 5m 15s content_loss:0.323784 style_loss:0.478477 total_loss:0.803261
3000 30.0% 7m 55s content_loss:0.282357 style_loss:0.476613 total_loss:0.758970
4000 40.0% 10m 36s content_loss:0.282233 style_loss:0.476295 total_loss:0.758518
5000 50.0% 13m 16s content_loss:0.282236 style_loss:0.476282 total_loss:0.758517
6000 60.0% 15m 57s content_loss:0.282230 style_loss:0.476288 total_loss:0.758518
7000 70.0% 18m 38s content_loss:0.282242 style_loss:0.476276 total_loss:0.758519
8000 80.0% 21m 19s content_loss:0.282240 style_loss:0.476277 total_loss:0.758518
9000 90.0% 24m 0s content_loss:0.282242 style_loss:0.476280 total_loss:0.759522
10000 100.0% 26m 41s content_loss:0.282245 style_loss:0.476272 total_loss:0.758517

torch.Size([1, 1, 257, 810])
torch.Size([1, 1, 257, 810])
1000 5.0% 2m 37s content_loss:0.116021 style_loss:0.008781 total_loss:0.125802
2000 10.0% 5m 19s content_loss:0.035642 style_loss:0.003484 total_loss:0.039126
3000 15.0% 7m 59s content_loss:0.011905 style_loss:0.002644 total_loss:0.014549
4000 20.0% 10m 41s content_loss:0.008837 style_loss:0.002632 total_loss:0.011470
5000 25.0% 13m 23s content_loss:0.008645 style_loss:0.002632 total_loss:0.011277
6000 30.0% 16m 4s content_loss:0.008643 style_loss:0.002632 total_loss:0.011275
7000 35.0% 18m 45s content_loss:0.008642 style_loss:0.002632 total_loss:0.011275
8000 40.0% 21m 26s content_loss:0.008646 style_loss:0.002629 total_loss:0.011275
9000 45.0% 24m 7s content_loss:0.008645 style_loss:0.002630 total_loss:0.011275
10000 50.0% 26m 48s content_loss:0.008637 style_loss:0.002637 total_loss:0.011275
11000 55.00000000000001% 29m 29s content_loss:0.008644 style_loss:0.002631 total_loss:0.011275
12000 60.0% 32m 10s content_loss:0.008643 style_loss:0.002631 total_loss:0.011274
13000 65.0% 34m 51s content_loss:0.008642 style_loss:0.002632 total_loss:0.011274
14000 70.0% 37m 33s content_loss:0.008643 style_loss:0.002632 total_loss:0.011275

```

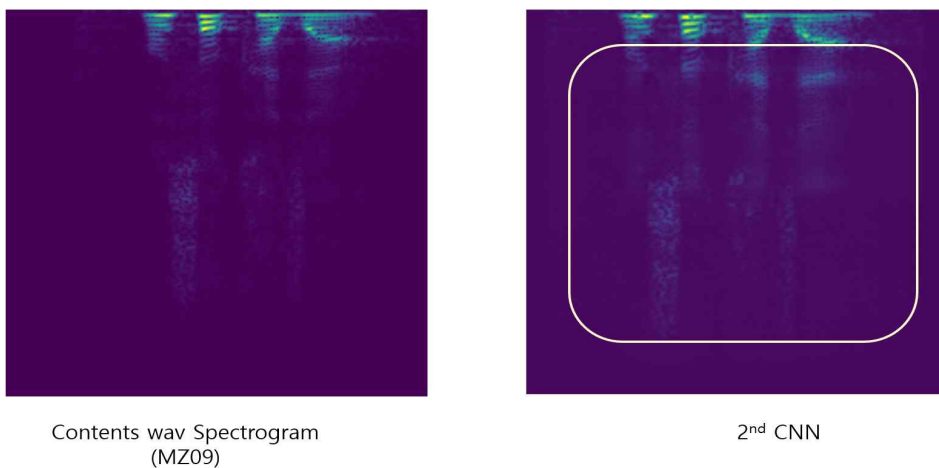
<그림 5-2> SeniorCNN의 훈련과정

## 5.4 SeniorCNN의 훈련결과

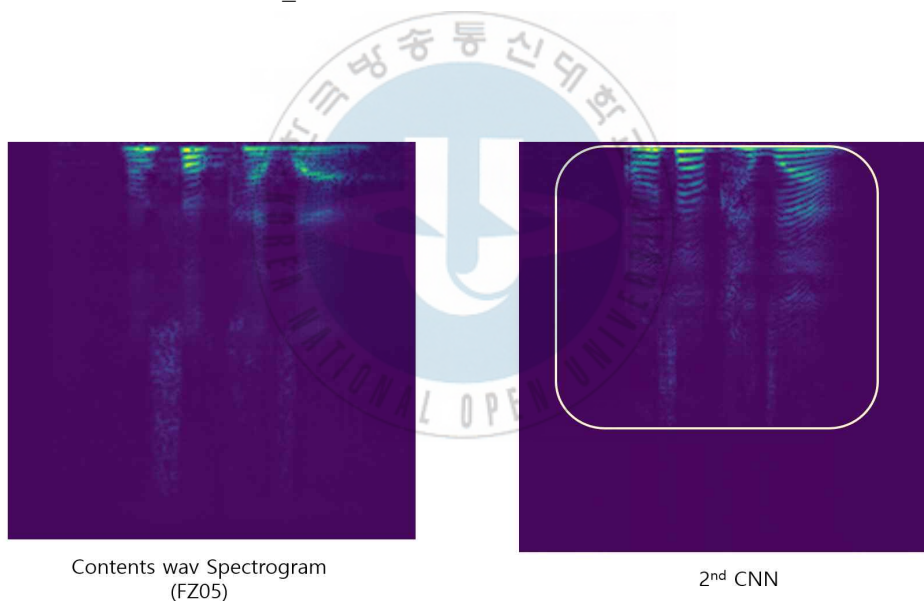
SeniorCNN 훈련결과, 원 데이터의 스펙트로그램 대비 SeniorCNN 훈련을 두 차례 수행한 음성 데이터의 스펙트로그램이 훨씬 선명하고 정교해짐을 확인할 수 있었다. <그림 5-3>과 <그림 5-4>는 “어서 가자”라는 문장을 발화한 음성이 원음성과 SeniorCNN 전 과정의 훈련을 거친 음성의 스펙트로그램이 어떻게 변화했는지를 보여준다.

<그림 5-3>은 71세의 남성화자 MZ09, <그림 5-4>는 68세의 여성화자 FZ05의 음성 데이터 변화를 나타내는 스펙트로그램이다. 스펙트로그램 상으로는 남성 화자보다는 여성 화자의 음성이 SeniorCNN 훈련 진행 후 더욱 정교하게 음성 데이터값을 갖게 되었다는 것으로 판단할 수 있다.





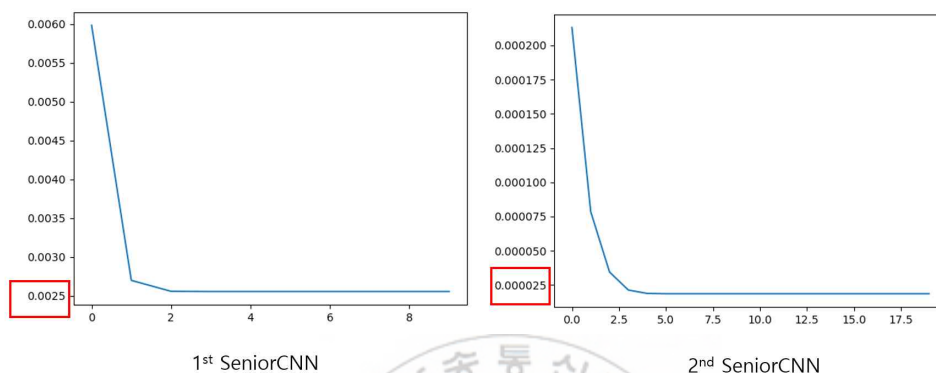
<그림 5-3> MZ\_09의 SeniorCNN 훈련 후 스펙트로그램의 변화



<그림 5-4> FZ\_05의 SeniorCNN 훈련 후 스펙트로그램의 변화

이러한 스펙트로그램의 변화는 활성화 함수 ReLU, Leaky ReLU, ELU 모두에서 같은 추세로 나타났다. SeniorCNN 훈련을 진행할수록 흐릿했던 스펙트로그램이 선명하게 되는 것을 활성 함수별로 출력한 음성 스펙트로그램을 통해서 확인할 수 있었다.

<그림 5-5>는 1<sup>st</sup> CNN과 2<sup>nd</sup> CNN을 거쳤을 때의 손실함수(loss function) 값의 변화량을 측정한 것이다. 손실함수의 경우, 2<sup>nd</sup> CNN에서 1<sup>st</sup> CNN의 결과 음성을 받아 훈련을 시작하므로 시작 지점의 손실함수 값이 훨씬 작다는 것을 확인할 수 있다.



<그림 5-5> 1<sup>st</sup> CNN과 2<sup>nd</sup> CNN의 손실함수 값의 변화량

또한, SeniorCNN의 전체 훈련과정을 거치게 되면 그 결과 음성 데이터가 원 음성 대비 잡음이 증가했다고 느껴지는데 그 잡음은 원 음성 데이터의 발화정보가 강화되었다는 것을 포함한다. 잡음이 섞인 음성 데이터가 음성인식 등 음성 인터페이스 사용에 좋은 데이터는 아니지만 잡음이 추가되었다는 이유만으로 본 SeniorCNN 훈련의 효용성이 없다고 판단하기는 어렵다. 음성 인터페이스의 사용은 사람과 하는 대화가 아닌 기계와의 상호작용이기 때문이다.

## 제6장 결과 분석

### 6.1 단어 오류율 측정 결과

#### 6.1.1 원시 데이터의 단어 오류율

아무 처리 없이 볼륨만 증가시킨 고령 화자의 원음성과 SPRocket의 실행으로 얻은 스타일 음성과 SeniorCNN으로 훈련 시켜 얻은 오디오 스타일 변환시킨 결과 모두 ETRI API[31]에 인식시켜 음성인식 실험을 수행하였다. 단어 오류율(WER, Word Error Rate)은 음성인식의 정확도를 측정하는 업계 표준으로 본 실험에서는 ETRI API를 호출해 음성을 인식한 결과 중 잘못 인식한 단어의 수를 센 다음, 실제로 발화한 단어의 총 개수로 나눈다. 이렇게 구한 해당 값을 마지막으로 100을 곱하여 백분율 계산으로 WER를 계산하였다.

수식 (1)은 WER를 구하는 공식을 나타낸 것이다.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

$S$  : 대체된 단어의 개수

$D$  : 삭제된 단어의 개수

$I$  : 삽입된 단어의 개수

$C$  : 교정된 단어의 개수

$N$  : 참조 문장의 전체 단어의 개수( $N = S + D + C$ )

$S$ (Substitution)는 발화한 단어가 아닌 다른 단어로 발음한 단어의 개수,  $D$ (Deletion)는 발화한 내용에는 있으나 음성인식 결과에서 빠진 단어의 개수,  $I$ (Insertion)는 발화한 내용에는 없는데 음성인식 과정에서 새로 삽입된 단어의 개수를 의미한다. 교정

된 단어의 수  $C(\text{Correction})$ 는 원래 발화한 단어의 개수를 셀 때 필요하다.

ETRI API 사용을 위해서 음성 데이터는 모두 16,000Hz, mono 형식을 사용하였다. 원시 데이터가 모두 16,000Hz, mono 형식을 만족하지만, SeniorCNN 훈련을 수행하면 데이터 보강으로 인해 샘플링 레이트가 22,050Hz로 출력하도록 설계하였다. 이렇게 샘플링 레이트가 바뀐 데이터는 모두 원시 데이터와 같은 음성파일 형식으로 변환하여 음성인식을 수행하였다. ETRI API는 16,000Hz, mono 형식의 wav 파일만 인식할 수 있기 때문이다.

WER는 파이썬 프로그래밍을 이용하여 구현하였다.

<표 6-1>은 고령화자 원시 데이터의 WER을 계산한 결과이다. Speaker\_ID는 ID\_연령으로 읽으면 된다. MZ\_09\_71은 Speaker\_ID가 MZ\_09, 그의 연령이 71세임을 나타낸다.

<표 6-1> 고령화자 원시 데이터의 단어 오류율

Speaker_ID	WER (%)
MZ_09_71	14.56
FZ_05_68	10.25

원시 데이터의 WER 값이 10%를 넘어가 그 대조군으로 동일한 낭독내용을 젊은 화자의 음성으로 녹음한 녹음파일을 ETRI\_API에 인식시켜 보았다. <표 6-2>는 25세 남성화자 MV\_13과 23세 여성화자 FV\_01의 음성의 WER을 계산한 결과이다.

<표 6-2> 젊은 화자 원시 데이터의 단어 오류율

Speaker_ID	WER (%)
MV_13_25	9.87
FV_01_23	10.54

<표 6-2>의 WER에서 알 수 있듯이 젊은 화자의 낭랑한 음성도 WER가 10% 가깝다는 사실에서 ETRI\_API의 통상적인 한국어 인식률을 고려해서 고령 화자의 WER를 살펴보아야 하며 이 WER 수치를 절대적인 값으로 SeniorCNN의 성능을 평가하기는 어렵다.

## 6.1.2 스타일 음성의 단어 오류율

<표 6-3> 고령화자 스타일 데이터의 단어 오류율

Speaker_ID	WER (%)
MZ_09_71_VC	13.89
MZ_09_71_DIFF_VC	12.36
FZ_05_68_VC	15.23
FZ_05_68_DIFF_VC	15.90

<표 6-3>은 SPRocket의 수행 결과로 본 SeniorCNN에서 스타일 음성으로 사용한 VC, DIFF\_VC의 WER를 나타낸 것이다. 스타일로 사용한 음성은 MZ\_09의 결과(MZ\_09\_71\_DIFF\_VC)가 WER이 12.36%로 가장 낮았으며, 같은 방식으로 보코더 없이 한 음성 변환(FZ\_05\_68\_DIFF\_VC)의 결과이지만 FZ\_05의 WER이 15.90%로 가장 높았다. 이를 통해 SPRocket의 음성 변환 결과는 특정 조건에서 변환했을 때 WER의 높고 낮음을 일률적으로 판단할 수는 없고 음성 데이터의 속성에 따라서 WER가 달라지는 특성을 가진다는 것을 확인할 수 있다.

이하 SeniorCNN 훈련에 사용한 활성화 함수별 결과 음성 데이터의 WER를 정리하겠다.

## 6.1.3 활성화 함수별 단어 오류율

<표 6-4> 활성화 함수별 SeniorCNN 훈련결과의 단어 오류율

Speaker_ID	Step	ReLU WER (%)	Leaky ReLU WER (%)	ELU WER (%)
MZ_09_71	1 <sup>st</sup> SeniorCNN	14.46	14.66	13.22
	2 <sup>nd</sup> SeniorCNN	16.00	17.43	17.24
FZ_05_68	1 <sup>st</sup> SeniorCNN	8.14	12.64	7.76
	2 <sup>nd</sup> SeniorCNN	11.11	13.89	12.55

<표 6-4>는 활성화 함수별로 SeniorCNN 훈련을 수행한 음성 데이터의 WER를 계

산한 것이다. <표 6-4>의 결과를 통해 세 가지 종류의 활성화 함수 모두 SeniorCNN의 전 과정 훈련을 진행한 음성보다 첫 번째 단계의 훈련만을 진행한 음성이 더 낮은 WER를 가진다는 것을 확인할 수 있다.

<표 6-5> Leaky ReLU SeniorCNN의 음성인식 결과

No.	Raw Data	1st CNN	2nd CNN
52	나무꾼은 미리 준비해 온 옷을 선녀에게 주었어여	나무꾼은 미리 준비해 온 옷을 손녀에게 주었어여	나무꾼은 미리 준비해 온 옷을 손녀에게 주었어요
53	선녀는 나무꾼이 건네준 옷을 입고 하는 수 없이 나무꾼을 따라가서 살게 되었지요	선녀는 나무꾼이 건네주는 옷을 입고 할 수 없이 나무꾼을 따라가서 살게 되었지요	선녀는 나무꾼이 건네주는 어릴 때 하는 수 없이 나무꾼을 따라가서 살게 되었지요
71	사슴이 아기를 셋 낳을 때까지 날개옷을 늦추지 말라고 한 뜻은 세 아이를 두 팔려면 앓을 수가 없었기 때문이었구나	4승 아기를 생각날 때까지 날개옷을 내주지 말라고 한 뜻은 세 아이를 두 팔로는 안할 수가 없었기 때문이었구나	사슴이 아기를 새마을 때까지 날개옷을 내주지 말라고 한 뜻은 세 아이를 두 팔로는 안을 수가 없었기 때문이었구나
72	아 이를 어찌나	아 이를 어찌나	아 이를 어찌나
73	나무꾼은 가슴을 치며 후회하고 눈물을 흘렸지만 선녀와 두 아이들은 영문 나무꾼이 사는 나라로 돌아오지 않았다는 슬픈 옛날 이야기랍니다	나무꾼의 가슴을 치며 후회하고 눈물을 흘렸지만 석류와 두 아이들은 영농 나무꾼이 사는 나라로 돌아오지 않았다는 스마트 폰 옛날 이야기랍니다	나무꾼은 가슴을 치며 후회하게 눈물을 흘렸지만 선녀와 두 아이들은 영농 나무꾼이 사는 마을에 돌아오지 않았다는 슬픈 옛날 이야기랍니다

두 번째 훈련을 진행하면서 이미 스타일 변환된 음성에 또다시 젊은 화자를 목표 화자로 하여 변환한 음성을 스타일로 하는 오디오 스타일변환은 원시 데이터를 오염시키는 역효과가 발생하여 단어 인식률을 떨어뜨린다는 것을 확인할 수 있었다.

WER만으로는 SeniorCNN 전 훈련과정을 평가하기에 미흡하여 음성인식으로 추출된 대본을 확인해 보았다. 이는 WER가 가장 높은 Leaky ReLU를 활성화 함수로 하여 훈련 시킨 데이터로 비교해 보았다. 가장 성능이 낮은 결과로도 SeniorCNN 훈련과정을 통해 음성인식 결과의 변화 추이가 있는지를 확인하기 위해서 Leaky ReLU로 훈련 시킨 데이터를 사용하였다.

<표 6-5>는 SeniorCNN 훈련의 단계별로 인식한 문장을 5개 선별하여 그 인식결과를 살펴본 것이다. 음성인식에 사용한 데이터는 여성화자 FZ\_05이다. No. 은 발화에 사용한 문장 번호를 나타낸다. SeniorCNN의 훈련과정을 진행할수록 원시 화자의 음성에서는 올바르게 인식했던 단어가 오히려 틀리게 되는 경우도 나타나기는 하지만 71번 문장의 “세 아이를 두 팔로는 안을 수가 없었기 때문이었구나” 구문은 훈련이 진행되면서 올바르게 수정되는 것을 확인할 수 있다.

이를 통해 SeniorCNN 훈련이 WER 결과치만으로 의미 없는 것이 아니라 실험한 훈련과정에서 발생하는 잡음을 최대한 줄이고 화자의 음성을 강화할 수 있도록 개선한다면 SeniorCNN의 훈련결과만으로도 고령 화자의 음성이라 하더라도 음성 인식률의 향상을 기대할 수 있을 것이다.

## 6.2 PESQ-MOS 측정 결과

### 6.2.1 PESQ-MOS 매핑 함수

매핑이란 두 집합의 원소 간의 대응 관계를 말하며 여기에서 말하는 매핑 함수(mapping function)는 일대일 대응 함수이다. ITU-T P.862 규약[20]에 따르면 PESQ 값은 -0.5에서 4.5 사이에 위치하기 때문에 주관적인 평가 기법인 MOS(Mean Opinion Score)와는 혼용하여 사용할 수 없는 것이 원칙이다. PESQ는 단방향 통신에서 청취

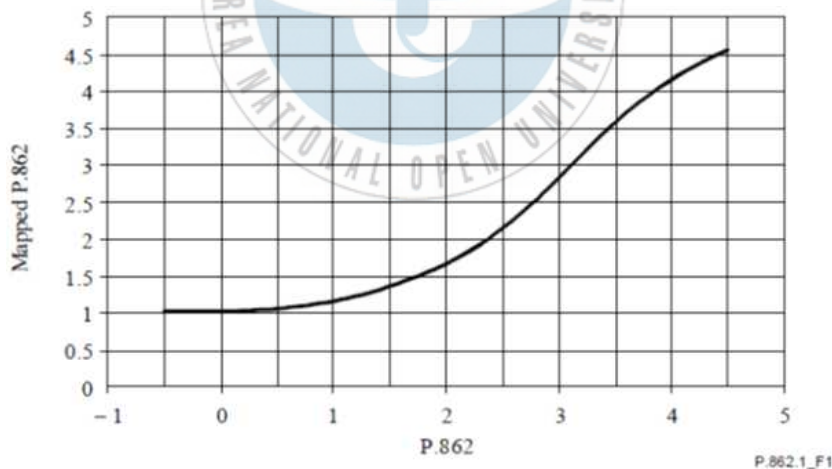


음성 왜곡 및 잡음을 측정할 수 있는 지표이다. 음향적으로 소리 세기의 손실, 지연, 에코 등의 전송 품질은 MOS와의 비교를 위하여 ITU-T P.862.1 규약으로 구한 PESQ 값으로부터 MOS-LQO(Listening Quality Objective, 이하 MOS라 한다)를 추정할 수 있는 매핑 함수를 제공한다.

이 함수로 인해 PESQ로부터 MOS를 산출할 수 있다. 아래의 수식 (2)는 PESQ 값으로부터 MOS값을 구하는 계산식이다. 수식 (2)에서  $x$ 는 원본 데이터와 합성음 데이터로 산출한 PESQ,  $y$ 는 수식 (2)로 계산되는 MOS 값이다.

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 * x + 4.6607}} \quad (2)$$

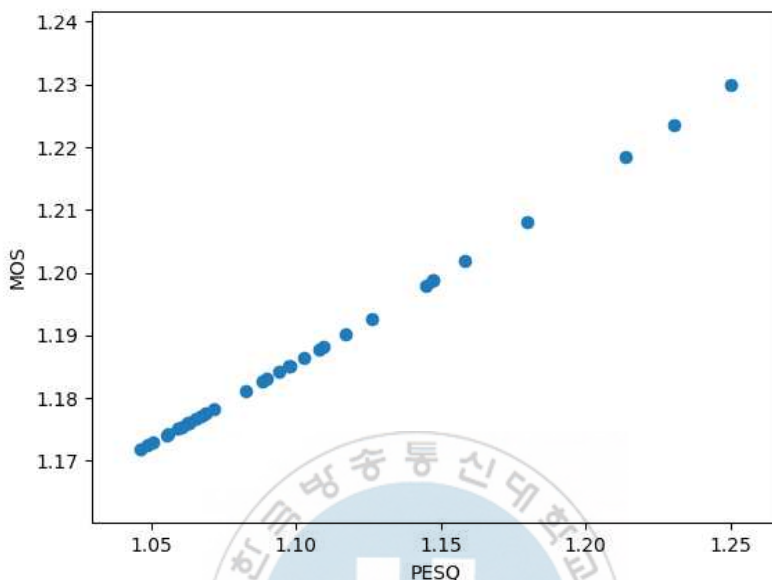
ITU-T P.862 규약에서는 ITU-T P.862.1의 PESQ 값으로부터 MOS-LQO를 추정할 수 있는 매핑 함수를 <그림 6-1>과 같이 제공한다. <그림 6-1>의 그래프는 식 (2)의 계산 결과로 구해지는 것이다.



<그림 6-1> PESQ 값으로부터 MOS를 추정할 수 있는 매핑 함수

<그림 6-2>는 실험에서 구한 음성 데이터별 PESQ에서 MOS를 구한 데이터로 산점도를 그린 것이다. 수식 (2)와 <그림 6-2>에서 확인할 수 있듯이 PESQ가 높으면 그

에 비례하여 높은 MOS가 산출된다.



<그림 6-2> PESQ-MOS의 산점도

## 6.2.2 원시 화자의 음성과 스타일 음성의 PESQ-MOS

PESQ 값은 참조 데이터가 존재해야 측정 가능하므로 MZ\_09, FZ\_05의 원시 음성을 참조 데이터로 사용하였다. PESQ 값은 참고문헌 [32-33]의 구현 내용을 기반으로 재구성한 파이썬 프로그래밍으로 구하였으며 이렇게 구한 PESQ 값으로부터 수식 (2)를 기반으로 MOS를 계산하였다. 화자별 33개 음성의 PESQ, MOS의 평균값은 <표 6-6>, <표 6-7>과 같다. <표 6-6>과 <표 6-7>의 상위 두 개의 행이 스타일 음성의 평균 PESQ, 평균 MOS를 계산한 값이다.

<표 6-6> MZ\_09의 PESQ, MOS 평균값

Speech Data	Mean PESQ	Mean MOS
MZ_09_71_VC	1.1544	1.2010
MZ_09_71_DIFF_VC	1.4151	1.2923
MZ_09_71_1st_CNN	1.8411	1.5295
MZ_09_71_2nd_CNN	1.5088	1.3344

<표 6-7> FZ\_05의 PESQ, MOS 평균값

Speech Data	Mean PESQ	Mean MOS
FZ_05_68_VC	1.0404	1.1705
FZ_05_68_DIFF_VC	1.3464	1.2645
FZ_05_68_1st_CNN	1.5737	1.3706
FZ_05_68_2nd_CNN	1.5027	1.3302

즉, ‘Speech Data’ 열의 MZ\_09\_71\_VC, MZ\_09\_71\_DIFF\_VC, FZ\_05\_68\_VC와 FZ\_05\_68\_DIFF\_VC는 오디오 스타일변환에서 스타일 음성으로 사용한 SPRocket의 결과 음성이다.

71세 MZ\_09 화자의 경우 SeniorCNN 1단계 훈련과정만을 통과한 음성 데이터가 평균 PESQ와 평균 MOS값 모두 네 개의 데이터 중 가장 높은 것으로 나타났다.

<표 6-7>의 FZ\_05 화자의 PESQ, MOS 평균값 역시 SeniorCNN 1단계 훈련과정만을 통과한 음성 데이터가 평균 PESQ와 평균 MOS값 모두 네 개의 데이터 중 가장 높은 것으로 나타나 객관적인 음성품질 평가지표에서는 SeniorCNN 1단계 훈련과정만을 통과한 음성 데이터가 가장 좋은 음질을 가진다는 것을 확인할 수 있다.

또한, 남성 데이터인 MZ\_71이 여성 데이터인 FZ\_05보다 수치상으로 우수한 음질을 가진다는 것을 알 수 있다.

### 6.2.3 활성화 함수별 PESQ-MOS

PESQ-MOS 역시 활성화 함수별로 그 값을 구하였다. <표 6-8> ReLU, <표 6-9>는 Leaky ReLU, <표 6-10>은 ELU를 활성화 함수로 하여 훈련을 한 결과 음성의 PESQ, MOS 평균값을 구한 것이다.

<표 6-8> ReLU로 훈련 시킨 음성의 PESQ, MOS 평균값

Speaker_ID	Step	Mean PESQ	Mean MOS
MZ_09_71	1 <sup>st</sup> SeniorCNN	1.8846	1.5645
	2 <sup>nd</sup> SeniorCNN	1.4998	1.3297
FZ_05_68	1 <sup>st</sup> SeniorCNN	2.0514	1.7008
	2 <sup>nd</sup> SeniorCNN	1.5528	1.3563

<표 6-9> Leaky ReLU로 훈련 시킨 음성의 PESQ, MOS 평균값

Speaker_ID	Step	Mean PESQ	Mean MOS
MZ_09_71	1 <sup>st</sup> SeniorCNN	1.8411	1.5295
	2 <sup>nd</sup> SeniorCNN	1.5088	1.3344
FZ_05_68	1 <sup>st</sup> SeniorCNN	1.5737	1.3706
	2 <sup>nd</sup> SeniorCNN	1.5027	1.3302

<표 6-10> ELU로 훈련 시킨 음성의 PESQ, MOS 평균값

Speaker_ID	Step	Mean PESQ	Mean MOS
MZ_09_71	1 <sup>st</sup> SeniorCNN	1.9046	1.5844
	2 <sup>nd</sup> SeniorCNN	1.4918	1.3259
FZ_05_68	1 <sup>st</sup> SeniorCNN	2.1126	1.7614
	2 <sup>nd</sup> SeniorCNN	1.5643	1.3606

<표 6-8>, <표 6-9> 및 <표 6-10>의 수치 데이터에서 확인할 수 있듯이 SeniorCNN 1단계만 훈련한 결과 음성이 2단계까지 훈련한 음성에 비해 그 음질이 좋았으며 활성화 함수로 ELU를 선택하여 훈련한 결과 음성이 남성화자 MZ\_09의 경우 평균 PESQ가 1.0946, 평균 MOS 수치가 1.5844로 가장 높았다.

여성 역시 활성화 함수를 ELU로 하여 훈련한 결과가 평균 PESQ 2.1126, 평균 MOS 1.7614로 가장 높은 결과를 나타내었다. 이러한 결과를 통해 활성화 함수 ELU가 SeniorCNN 훈련에서 가장 우월한 결과를 나타내는 활성화 함수임을 확인할 수 있었다.

WER, PESQ-MOS 수치를 종합적으로 분석한 결과 활성화 함수  $ELU > ReLU > Leaky ReLU$  순서로 결과 음성의 음질이 떨어지는 것을 확인할 수 있었다.



## 제7장 결론

본 연구는 고령 화자의 음성이 갖는 취약점을 보완하기 위해 진행하였다. 이를 위해 오디오 스타일변환 기법을 사용하였고, 이는 오디오 파일에 적용하기에는 잡음이 너무 많다는 한계점을 고려하여 GMM 기반 음성 통계학적인 기법을 사용하는 SPRocket이라는 음성 변환 소프트웨어를 사용하여 스타일 음성으로 사용할 데이터를 추출하였다. SPRocket으로 추출한 음성을 스타일변환의 스타일 음성으로 사용하는 것이 다른 젊은 화자의 음성을 스타일 음성으로 사용한 것보다 훨씬 잡음이 감소하는 것을 확인할 수 있었다.

또한, SeniorCNN이라는 고령화자 음성 강화 및 복원에 사용하는 딥러닝 모델을 제안하였다. CNN의 훈련을 위해 어떠한 활성화 함수가 가장 좋은 성능을 내는지를 비교하기 위해 대표적인 활성화 함수 세 가지, ReLU, Leaky ReLU 그리고 ELU 함수를 사용하여 SeniorCNN을 각각 훈련 시켰다. 그 결과 남성과 여성 음성 모두 ELU 함수로 훈련한 음성의 WER, PESQ-MOS 결과가 가장 높은 결과값을 나타냄을 확인할 수 있었다.

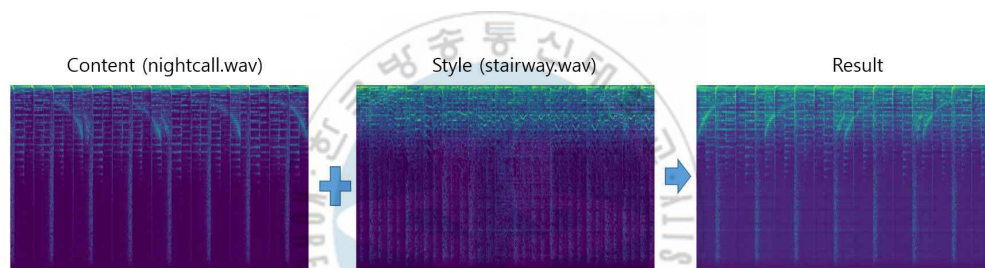
SeniorCNN 훈련을 한 번으로 끝낸 것이 아니라 2단계로 나누어 수행한 것도 의미가 있었다. 훈련을 여러 단계 오랜 시간 지속한다고 하여 결과 음성의 품질이 반드시 좋아지지 않는다는 것을 2단계 실험을 통해 다시금 확인할 수 있었다.

이에 더하여 본 연구에서는 음성의 스타일을 변환한 결과의 음질을 객관적으로 분석하는 시도를 하였다. 음성인식의 성능 지표로 대표적으로 사용되는 WER뿐만 아니라 ITU-T 표준 규약인 PESQ를 SeniorCNN 훈련으로 새로 생성한 음성의 품질 평가에 활용하였다.

음성합성, 음성 변환 등의 연구 분야는 그 결과물에 대한 평가가 주관적으로 설문조사 등의 기법을 통해 수행해 왔기 때문에 평가자가 누구냐에 따라 그 결과가 일정하지 않았고 그에 들이게 되는 비용도 또한 매우 높았다. 그런 점에서 본 시도는 음성 연구의 결과물을 객관적으로 분석하게 하면서도 비용이 거의 들지 않았다는 점에서 종래의 음성연구 결과의 평가방법을 개선하였다.

SeniorCNN은 고령 화자의 음성 데이터 보완을 위해 고안한 것이기는 하나 이는 사람의 음성 데이터뿐만 아니라 음악 데이터에도 활용 가능성이 있다. 음악 파일 처리에 관련하여 음악 데이터의 스타일변환에 활용할 수 있다는 사실을 SeniorCNN 훈련을 통해 확인하였다. <그림 7-1>은 두 개의 연주곡을 하나의 음악 파일로 만든 것이다. 스타일 변환시킨 음악(Content)은 기계로 만든 반주 음원이고, 스타일로 입힌 음악은 기타 연주 음악이다.

오디오 스타일변환 선행 연구[18-19]에서 음악을 스타일 변환한 이후 결과가 무슨 음악인지 알아듣기 어려웠던 것과는 달리 SeniorCNN을 이용해 훈련한 결과는 두 장르의 음악이 모두 하나의 결과 파일에 남아 있으면서 듣기에 편안한 음원을 생성해 낼 수 있었다.



<그림 7-1> SeniorCNN으로 구현한 음악의 스타일변환

SeniorCNN에 잡음 제거 능력을 보완한다면 반주 없는 독창이나 독주에 반주를 입히는 응용 등 음악적으로 복수 개의 음원을 합쳐야 할 필요성이 있을 때 유용하게 활용할 수 있으리라 생각한다.

또한, 현재 구현한 SeniorCNN의 단점을 보완한다면 어학교육 분야에서도 외국어로 언어를 배우는 화자들의 발음교정에 SeniorCNN이 사용될 수 있다고 생각한다. 그러면 성인이 되어 외국어를 배운다고 해도 모국어가 아닌 외국어를 학습하면서 가장 어려운 과제 중 하나인 발음을 정확하게 구사하지 못하는 문제를 해결하는 데에 도움을 줄 수 있을 것이다.

SeniorCNN은 일반적인 보급형 PC로는 GPU가 장착되어 있더라도 단시간 내에 그 결과를 볼 수 있는 모델은 아니다. 스타일 음성을 SPRocket을 통해 추출하는 것과



SeniorCNN 훈련단계를 거치는 시간이 상당히 소요되기 때문이다.

또한, SeniorCNN은 병렬 말뭉치 기반으로 개발되었다. 알고리즘의 효율성을 생각한다면 비병렬 말뭉치를 이용해 원시 화자와 목표 화자가 발화한 내용이 달라도 음성 변환이 가능하도록 구현하는 것이 훨씬 향후 기술의 활용범위가 넓지만, 결과 음성의 품질 문제와 하드웨어 자원의 한계로 비병렬 말뭉치의 음성 스타일변환까지는 구현하지 못하였다. 현재 구현한 SeniorCNN은 음성 변환하고자 하는 원시 데이터와 스타일 데이터가 대본이 모두 동일한 내용을 발화한 것이어야 사용할 수 있다.

또한, 음성의 스타일을 변환하면서 생기는 잡음 문제도 해결해야 할 과제이다. 이미 지나 7장에서 응용 분야로 제시한 음악에 SeniorCNN을 적용한다면 잡음도 새로운 시각적·음향적인 효과를 낼 수도 있어 큰 문제가 되지 않겠지만 현재의 기술 수준으로는 잡음이 음성 인터페이스를 사용하는데 큰 장애 요소이므로 반드시 극복해야 할 대상이다. 한국인들이 개발해 한국어의 음성 인식률이 높은 축에 속하는 ETRI의 음성인식 API로 음성인식 작업을 수행해도 단어 오류율이 보통 10%대가 나오기 때문에 SeniorCNN에서 잡음 문제는 묵과할 수 없는 문제이다.

이러한 한계점에도 불구하고 본 연구는 저출산·고령화의 심각한 문제점을 안고 있는 한국 사회에 필요한 연구이고 SeniorCNN이 갖는 단점을 보완할 수 있다면 그 활용범위는 무궁무진할 것이다.

한국은 이미 고령사회가 되었고 초고령화 사회로의 진입을 앞둔 시점에서 고령 화자의 음성 데이터에 관한 연구는 그 수요와 공급 모두의 측면에서 필요하다. 특히 고령 화자의 음성 자체가 가지고 있는 약한 발성과 성대 떨림, 많은 잡음 등의 문제를 해결해 기계에서도 음성인식 및 음성 데이터 전송이 자유자재로 이루어질 수 있도록 하는 연구가 필요하다. 향후 맞춤형 건강관리 및 스마트 홈케어 시스템에도 노인 음성의 특성을 반영해 적극적으로 개발하고자 하는 법적, 제도적 장치가 마련되어야 할 것이다.

## 참고문헌

- [1] Rachel Batish, “보이스 & 챗봇 디자인”, 에이콘출판주식회사, 2019년.
- [2] 김동성, “언어지능”, 커뮤니케이션북스(주), 2017년.
- [3] 염태정. 65세 이상 인구 첫 14% 넘어…한국 '고령사회' 첫 공식 진입.  
중앙일보. 2017년 09월 03일.
- [4] The Body Odd, (2013, Jan. 26), The wavery, shaky 'old person's voice,' explained [Online].  
Available: <https://www.nbcnews.com/health/body-odd/wavery-shaky-old-persons-voice-explained-f1C8119298> (downloaded 2019, Oct. 3)
- [5] Soonil Kwon, Sung-Jae Kim, Joon Yeon Choeh, “Preprocessing for elderly speech recognition of smart devices,” Computer Speech & Language, Vol. 36, pp. 110-121, Mar. 2016.
- [6] G. Son, S. Kwon, and Y. Lim, “Speech rate control for improving elderly speech recognition of smart devices,” Advances in Electrical and Computer Engineering, Vol. 17, No. 2, pp. 79-84, 2017.
- [7] 이윤주, 오세영, 이순규, 배명진, “HMM을 이용한 음성인식 시스템의 전처리에 관한 연구,” 대한전자공학회 1999년도 추계종합학술대회 논문집, pp. 668-671, 1999년 11월.
- [8] Mary Jo Creaney-Stockton. (1996, August). Chapter 3 Preprocessing Of The Speech Data [Online].  
<https://my.fit.edu/~vkepuska/ece5526/Chapter%203%20Preprocessing%20Of%20The%20Speech%20Data.htm> (downloaded 2019, Oct. 15)

- [9] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, , B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.
- [10] 오타 미즈하사, 수도 코다이, 쿠로시와 타쿠마, 오다 다이ске, "실전! 딥러닝-텐서플로와 케라스를 이용한 딥러닝 최신 기술 활용 가이드," 위키북스, 2019년.
- [11] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss," Proceedings of the 36th International Conference on Machine Learning, in PMLR 97, pp. 5210-5219, 2019.
- [12] T. Kaneko, H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," arXiv preprint arXiv:1711.11293, 2017.
- [13] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGANVC: Non-parallel many-to-many voice conversion with star generative adversarial networks," arXiv:1806.02169 [cs.SD], June 2018.
- [14] K. Kobayashi, T. Toda, "sprocket: Open-Source Voice Conversion Software," In Odyssey (pp. 203-210), June, 2018.
- [15] K. Kobayashi. (2018). SPRocket system [Online].  
Available: <https://github.com/k2kobayashi/sprocket> (downloaded 2019, Sep. 5)
- [16] Tomoki TODA. (2018). Hands on Voice Conversion. [Online].  
Available: [https://www.slideshare.net/NU\\_I\\_TODALAB/hands-on-voice-conversion](https://www.slideshare.net/NU_I_TODALAB/hands-on-voice-conversion) (downloaded 2019, Sep. 5)

- [17] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414-2423, 2016.
- [18] E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez, "Audio style transfer," In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 586-590, IEEE, April, 2018.
- [19] Alish Dipani. (2018). (INTEL AI DEVELOPMENT PROGRAM) Neural Style Transfer on Audio Signals. [Online].  
Available: <https://software.intel.com/en-us/articles/neural-style-transfer-on-audio-signals> (downloaded 2019, Sep. 25)
- [20] ITU. (2001). P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs  
Available: <https://www.itu.int/rec/T-REC-P.862> (downloaded 2019, Sep. 15)
- [21] 임창송, 배건성, "HMM 기반의 한국어 합성음에 대한 PESQ 및 MOS 평가의 상관도 분석," 말소리와 음성과학, 2권, 1호, 71-75쪽, 2010년
- [22] 장대익, "노인 음성에 대한 노인음성지수와 기기적, 청지각적 평가 간 상관성," 이학석사(언어치료)논문, 대구대학교, 2019년.
- [23] P. Pontes, A. Brasolotto, and M. Behlau, "Glottic characteristics and voice complaint in the elderly," Journal of Voice, Vol. 19, No. 1, pp. 84-94, 2005.
- [24] 김선해, 고도흥, "한국 정상 노인음성의 기본 주파수," 음성과학, 15권, 3호, 95-102쪽, 2008.
- [25] 김진중, "골빈해커의 3분 딥러닝 텐서플로맛," 한빛미디어(주), 2017년.

- [26] 라온피플. (2016년 1월 25일). CNN의 구조. 라온피플 머신러닝 아카데미 [온라인].  
사용 가능:  
<http://blog.naver.com/PostView.nhn?blogId=laonple&logNo=220608018546&categoryNo=0&parentCategoryNo=0&viewDate=&currentPage=1&postListTopCurrentPage=1&from=postView> (downloaded 2019, Oct. 5)
- [27] Mazzy Star. (2018). Voice style transfer with random CNN [Online].  
Available: <https://github.com/mazzystar/randomCNN-voice-transfer> (downloaded 2019, Sep. 30)
- [28] 분석왕 분석벌레. (2019년 6월 3일). [신경망] 6. 활성화 함수 (Activation Function) [온라인].  
사용가능: <https://analysisbugs.tistory.com/55> (downloaded 2019, Oct. 1)
- [29] 장지수, “딥러닝에 목마른 사람들을 위한 PyTorch-개인용 GPU 학습 서버 구축부터 딥러닝까지,” 비제이퍼블릭, 2019년.
- [30] Fei-Fei Li, Justin Johnson, and Serena Yeung. (2017). Stanford University cs231n Lecture Note 6 [Online].  
Available: [http://cs231n.stanford.edu/slides/2017/cs231n\\_2017\\_lecture6.pdf](http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture6.pdf) (downloaded 2019, Oct. 21)
- [31] ETRI. (2017). ETRI OpenAPI [Online].  
Available:[http://aiopen.etri.re.kr/guide\\_recognition.php#group01](http://aiopen.etri.re.kr/guide_recognition.php#group01) (downloaded 2019, Mar. 1)
- [32] imankulov. (2009). Python functions to convert between different speech quality metrics [Online].  
Available: <https://github.com/imankulov/vqmetrics> (downloaded 2019, Sep. 15)

- [33] ludlows. (2019, May). PESQ (Perceptual Evaluation of Speech Quality) Wrapper for Python Users (narrow band and wide band) [Online].

Available: <https://github.com/ludlows/python-pesq> (downloaded 2019, Sep. 15)



# A B S T R A C T

## A Study for the elderly Speech Data Preprocessing Technique with Voice Conversion based on Audio Style Transfer

by

***Woo Eun Ju***

Department of Computer Science  
Graduate School  
Korea National Open University

Supervised by Professor : Lee. Byeong Rae

The Conversational User Interface (CUI) is a new paradigm that interacts with computers by mimicking the natural conversations of people as a result of long-standing human-machine interactions. This technology is especially for elderly people who are not familiar with the operation of electronic devices and cannot use them easily. In 2018, more than 14% of the Korean population aged 65 or older: Korea entered full-scale aging society. In spite of these social phenomena, studies on the characteristics of aging voices and their improvement and restoration are hardly conducted in Korea. The aging voice is small, the pronunciation is inaccurate and it is not suitable for the



general speech recognizer made based on the standard pronunciation of the young. In consideration of the voice characteristics of the elderly speakers, this study improves the recognition performance of the speech recognition system used by elderly people who are uncomfortable and elderly speakers who suffer from senile voice diseases by preprocessing the voice data of the elderly speakers using audio style transfer. In order to improve the convenience of use, simple preprocessing effects, such as noise reduction and normalization, have been made to compensate for the shortcomings of the existing voice technology, which cannot generate new data that is not in the original data. In this study, experiments were carried out using a Seoul speech reading corpus supplied by the National Institute of Korean Language in 2005. Among the data, one elderly speaker with the lowest voice quality was selected per gender. 33 speech data of 68-year-old female speaker and 71-year-old male speaker were subjected to voice conversion. The style voice used for audio style transfer is the result of SPRocket developed by Nagoya University in Japan. The result of voice conversion was made by selecting source speaker as aging speaker and target speaker as young speaker in 20s.

The deep learning algorithm proposed in this study is SeniorCNN, which consists of training 2D Random CNN two times. Two style voices are required to operate SeniorCNN designed to apply the style of voice converted into SPRocket to voice of elderly voice. In the first step, general speech conversion results using GMM-based traditional speech statistical techniques were used. In the second step, vocoder-free-based speech conversion results were used as style speech.

This result data was recognized by ETRI's public speech recognition API to compare WER (Word Error Rate) for each voice conversion task step. We compared the results of SeniorCNN's speech with the quality of SPRocket's speech and the voice data itself, and the speech recognition performance of the data. WER increased by 3% overall when passing both steps of SeniorCNN, but the result of ETRI voice recognition does not always output a constant text every time whenever the speech data is transmitted and recognized by the API. Given the Just only WER, it is difficult to assess the absolute performance of SeniorCNN. For this reason, the results of the speech recognition showed that there are cases that the words and phrases that were not recognized in the voice of the old aged speaker before the SeniorCNN training were recognized as the correct words and phrases when they were recognized by ETRI API after SeniorCNN training.

As a result of analyzing the PESQ for each voice and the value converted to MOS, the value trained by SeniorCNN using the style voice was higher than that of SPRocket. However, I could confirm that the results of SeniorCNN training are superior to both WER and PESQ-MOS, compared to the intermediate CNN passing only the first CNN training to the second CNN. As a result of the experiments, repeating CNN training several times did not necessarily indicate good results.

In addition, this study compared and analyzed how the final result voice quality varies according to the activation function used in CNN training. The most commonly used activation functions, ReLU and ReLU-derived activation functions, Leaky ReLU were used. SeniorCNN training by using these three

activation functions showed that for both male and female speakers, the voice files trained using ELU had the lowest WER and the PESQ and MOS were the highest. This suggests that the activation function ELU is most suitable for SeniorCNN training for speech reinforcement and reconstruction of elderly speakers.



## 감사의 글

대학원에 입학한 지 엇그제 같은데 어느덧 2년 6개월이라는 시간이 흘렀습니다. 학업과 업무를 병행하는 대학원 생활이기에 쉽지 않을 것이라 예상은 했지만 지난 대학원 생활은 많은 것을 참아 온 시간이었습니다. 그 인내의 시간은 학부를 졸업한 지 오래되어 공부의 필요성을 느껴 회사 업무와 병행할 수 있는 교육기관에서 공부하고 싶었던 저의 선택이었기에 후회는 없습니다. 짧지 않았던, 그렇지만 할 일은 참 많았던 대학원 생활을 끝까지 지속할 수 있었던 건 가족들의 지원, 훌륭하신 교수님들의 지도와 함께 공부했던 원우님들의 격려가 있었기에 가능한 일이었던 것 같습니다.

먼저 이 논문을 위해 늦은 시간까지 연구실에서 저를 기다려 주시고 논문 내용은 물론 단어 하나까지 세심하게 살펴 주신 지도교수님, 이병래 교수님께 감사의 말씀을 드립니다. 또한, 심사 과정에서 꼼꼼히 살펴 주시고 제 논문을 완성하는 데 많은 조언을 해 주신 정재화 교수님과 김강현 교수님께 감사의 말씀 올립니다. 그리고 대학원 정보과학과에서 수학했던 시간 동안 좋은 강의로 많은 가르침을 주셨던 교수님들께도 감사의 말씀 올립니다.

즐거지만은 않았던 대학원 생활의 활력이 되었던 한양 스터디 선배, 후배님들 모두 이 논문이 나오기까지 아낌없는 조언과 격려 보내 주신 점 너무 감사드립니다. 그리고 정보과학과 원우님들 전원 학업을 도중에 포기하지 마시고 석사학위를 꼭 취득하시기를 소망합니다. 또한, 대학원 생활의 고충을 이해하고 응원해 주신 회사 직원분들께도 감사의 말씀을 전합니다.

마지막으로 대학원 공부를 묵묵히 응원해 주셨던 사랑하는 엄마, 아빠께 항상 감사드립니다. 그리고 늘 열심히 사는 오빠에게 고맙다는 인사를 전하며 제 감사의 글을 이만 줄이겠습니다.

## 학위논문 공개 동의서

한국방송통신대학교 대학원은 귀하의 학위논문을 데이터베이스화하여 내용의 일부 또는 전부를 어떤 형태로든 국내·외 이용자에게 공개(열람 또는 배포)하는 것에 대하여 동의를 얻고자 하오니 동의여부를 기재하여 주시기 바랍니다.

학 위	이학석사	졸업년도	2020
전 공	정보과학과		
학위논문 제 목	한글	오디오 스타일변환 기반 음성 변환을 이용한 고령 화자의 음성 데이터 전처리 기술에 관한 연구	
	영문	A Study for the elderly Speech Data Preprocessing Technique with Voice Conversion based on Audio Style Transfer	
동 의 여 부	1. 찬성( <input checked="" type="radio"/> )    2. 조건부 찬성( <input type="radio"/> )    3. 반대( <input type="radio"/> )		
조건 또는 사유	※ “조건부 찬성” 또는 “반대”하는 경우 그 사유 기재		

2019 년    01 월    20 일

논문저자명 : 우 은 주



(서명 또는 날인)

**한국방송통신대학교총장 귀하**