■ 연구과제 요약문

과제명(기간)	딥러닝을 활용한 화자 특성 반영 및 사용자 맞춤형 TTS 기발 기술 개발 (2017-04-01~2019-03-31)
연구책임자	이 교 구
개요	- 본 연구 과제의 전반부에서는 사용자 맞춤형 TTS 기반 기술 개발을 위한 요소기술 중 하나인 오디오-문자열 동기화에 대한 시스템 개발 및 검증 과정을 진행하였음 - 특히, 가사와 음원 쌍을 항상 얻을 수 있는 음악 음원에 대한 가사동기화를 주제로 하였으며, 이를 통해 추후 개발할 TTS 시스템에 대한 데이터 전처리를 용이하게 할 수 있도록 하였음. - 가사동기화: 현재 멜론등 스트리밍 서비스에서는 일반적으로 음악오디오와 가사사이의 동기화 및 프롬프터 기능을 제공하고 있으며, 해당 동기화 과정은 외주업체를 통해 진행되고 있으며 사람이 직접 만든 데이터를 활용하고 있음.  그림 1 가사동기화 예시 - 원활한 서비스의 제공을 위해서 신곡에 대한 가사와 오디오의 정렬이 필수적이지만, 이는 많은 시간과 비용, 인력이 투입되는 과정이므로 이를 자동화했을 때 이익이 발생할 수 있음. - 또한 해당 연구는 자동 가사동기화 시스템을 설계할 수 있다는 점을 넘어서 오디오와 텍스트 사이의 연관성을 모델링할 수 있다는 점에서 향후 TTS system, source separation 등 다양한 음성 관련 Task에 적용할 수 있는 기술이라는 점에서 의의가 있음.
연구개발 결과	- 가사동기화 문제는 크게 두 가지 측면에서 도전적인 과제이다. 첫째로, 일반적인 음원의 경우 악기 등의 배경음악과 보컬의 가창이 함께 섞인 채로 포함되어 있기 때문에 가사동기화에서 필요로하는 깨끗한 보컬 음원만을 분리해낼 수 있다는 문제가 있다. 둘째로 해당 음원을 얻을 수 있는 경우에도 이를 문자 정보와 정렬하는 것 역시 서로 다른 두 도메인의 신호를 연관지어야 한다는 점에서 어렵다. - 위와 같은 문제를 해결하기 위하여 우리는 악기 및 배경음악이 포함된 혼합 음원 속에서도 보컬의 가창 음원만을 문자로 전사하는 가사동기화 모델을 개발하였다. - 또한, 해당 가사동기화 모델을 사용하여 추정된 가사와 실제 그 노래의 가사 사이를 동기화시키는 다이나믹-프로그래밍 기반 알고리즘을 개발하였다.

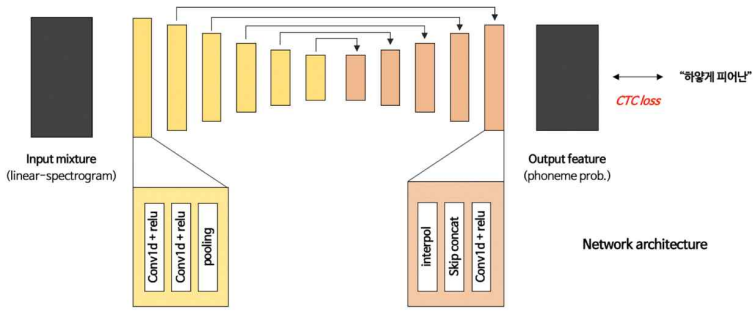


그림 2 가사동기화 인공지능 모델

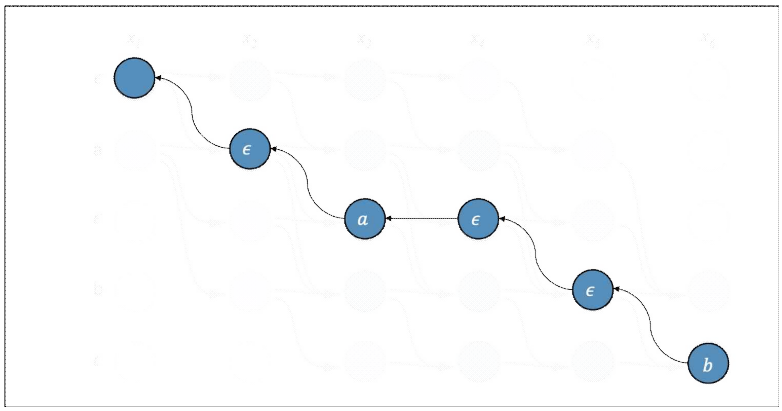


그림 3 Viterbi 알고리즘을 활용한 가사동기화 알고리즘

- 실험 결과 우리의 모델은 테스트 데이터셋에 대해서 실제 동기화 정보와 0.2초 이내의 오차만을 갖는 정확한 동기화를 진행할 수 있음을 보였다. 또한 우리가 제안한 알고리즘은 복잡한 배경음악이 섞여있는 경우에도 강인하게 동작한다는 특징을 갖는다.
- 또한 우리는 해당 결과를 활용하여 음원과 해당하는 가사를 입력했을 때 자동으로 송 프롬프터로 동작하는 간단한 소프트웨어를 제작했으며, 이를 통해 동기화가 정상적으로 이루어짐을 확인했다.

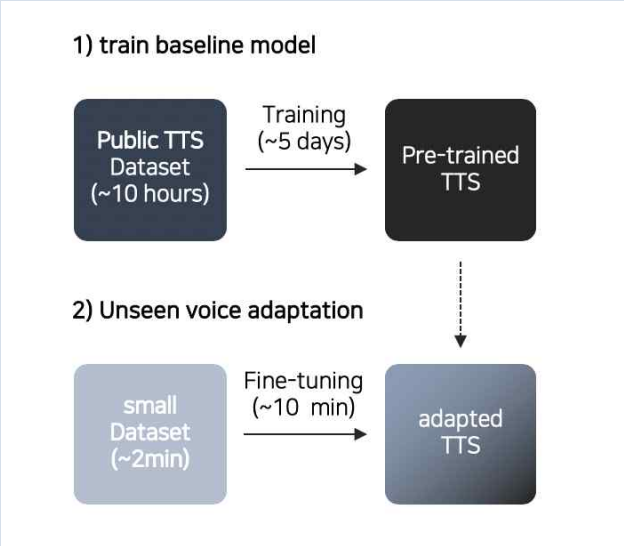
활용분야 및 기대효과

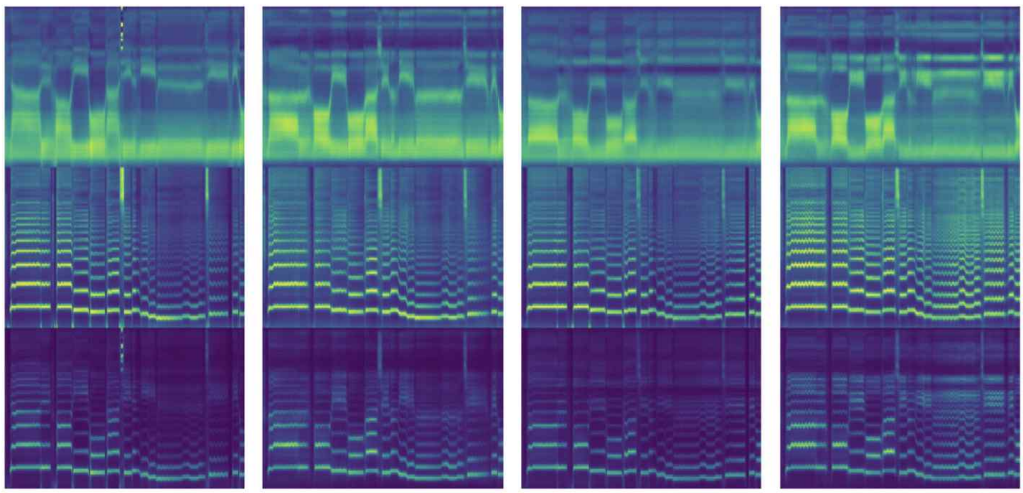
- 인공지능을 활용한 고성능의 가사동기화 모델이 효과적으로 훈련된다면 사람의 수작업이 없이도 가사동기화를 자동화하여 많은 비용을 절감할 수 있을 것으로 예상할 수 있다.

- 멜론의 플랫폼을 활용할 시 노래방 어플 등의 사업화 영역에 사용될 수 있는 가능성도 존재한다.

- 보다 학술적인 측면에서, 본 연구를 통해 오디오 정보로부터 그에 대응되는 문자열 정보 사이의 관계성을 찾아낸점을 활용하여 향후 오디오-텍스트 두 가지 영역에 대한 다양한 문제들 (음성 인식, 음성 합성 등) 에 도움을 줄 수 있는 기술로서 활용될 수 있다. (정확한 annotation을 포함한 데이터셋 제작, 문자열-오디오 정렬을 통한 추가적인 정보 입력 등)

■ 연구과제 요약문

과제명(기간)	딥러닝을 활용한 화자 특성 반영 및 사용자 맞춤형 TTS 기반 기술 개발 (2017-04-01~2019-03-31)
연구책임자	이 교 구
개요	- 본 연구 과제의 후반부에서는 사용자 맞춤형 TTS 기반 기술 개발을 위한 요소기술 중 하나인 소용량 음색 복제 및 합성에 대한 연구 초기 실험 및 개발을 진행하였음. - 딥러닝 기술의 발전으로 최근 전통적인 방식으로 진행되어오던 음성 합성 분야에 인공지능 기술이 성공적으로 적용되고 있으며, 실제 음성에 가까운 음질, 다양한 감정 표현등이 가능해지고 있음. - 그러나 데이터로부터 훈련되는 딥뉴럴네트워크의 특성상 하나의 목소리에 대한 모델을 만들기 위해서는 수시간~수십시간에 달하는 고품질 음성 데이터가 필요하다는 한계는 여전히 존재하며, 이러한 데이터 수집에 큰 시간과 비용이 소요되기 때문에 실제 시스템 적용에 진입장벽이 존재하는 상황임 - 이에 본 연구에서는 공개된 대용량 데이터셋으로 훈련한 모델을 바탕으로, 소용량의 새로운 목소리 샘플만을 가지고 음색을 복제해 해당 목소리로의 자연스러운 음성 합성이 가능한 음색 복제 기술 및 합성 시스템을 제안하고 분석하였음.  [소용량 음색 복제 TTS 시스템 훈련 모식도]
연구개발 결과	- 본 연구에서는 대표적인 딥러닝 기반 음성 합성 네트워크인 Tacotron을 기본 모델로 하여 이를 발전시킴으로써 소용량 복제가 가능한 다화자 음성 합성 네트워크를 설계 및 제안하였음. - 일반적인 음성합성 네트워크의 경우 특정 화자의 발화 (speech) 오디오와 그에 해당하는 문자열 (text) 정보를 훈련 데이터로 활용하며, 훈련된 모델은 주어진 임의의 입력 텍스트에 대해 그를 자연스럽게 발음한 speech를 생성해낼 수 있도록 훈련됨. 이 때, 발음 이외의 화자 특성 정보들 (억양, 음정, 빠르기 등)은 직접적으로 입력되지 않기 때문에, 네트워크는 스스로 이러한 화자와 관련된 여러 가지 특성을 학습해야 함

	- 본 연구에서 우리는 음정과 발음을 독립적으로 입력하는 방식으로 기존의 TTS 시스템을 확장함으로써, 소용량 데이터를 활용한 음색 복제 시 네트워크가 학습해야하는 요소를 줄이고 이로 인해 보다 효율적으로 적은 데이터에서 복제가 가능하도록 했음. - 실험 결과 제안한 모델은 약 20문장 (2-3분)의 새로운 화자의 음성 데이터가 있는 경우, 해당 화자의 음색 특징을 효과적으로 복원해낼 수 있음을 확인하였음. - 또한 새로운 화자의 음색을 학습하는데에는 총 10분 내외의 학습 시간만을 필요로 함을 확인하였으며, 이를 통해 소용량의 데이터로 빠른 시간 안에 해당 화자의 음색을 반영한 TTS 모델을 생성해낼 수 있게 되었음  [서로 다른 화자에 대해 같은 음정, 텍스트에 대해 생성된 spectrogram 결과, 위에서부터 formant mask, pitch skeleton, 그리고 mel spectrogrma. 음정은 유지한 채 formant mask만 변화하며 음색을 모델링하는 것을 확인할 수 있음]
활용분야 및 기대효과	- 본 연구에서 우리는 음정과 발음을 독립적으로 제어 가능한 TTS 시스템을 제안하였으며 이는 기존에 주어진 텍스트에 대해서만 적절한 음정과 속도의 음성을 생성해내는 네트워크와는 다르게 세밀한 조정이 가능하다는 점에서 의의가 있다. 이를 통해서 보다 풍부한 감정 표현, 원하는 억양 등을 반영한 음성을 합성할 수 있다. - 또한 제안한 모델은 두 요소를 독립/직접적으로 제어할 수 있기 때문에 새로운 목소리의 데이터에 빠르게 적응하는 것을 확인하였다. 새로운 화자의 2분정도 음성을 통해 해당 화자의 목소리를 생성해내는 TTS를 훈련시킬 수 있으며, 이는 기존 수 시간에서 수십분이 필요하던 시스템들에 비해 경제적인 비용으로 새로운 TTS 시스템을 만들어낼 수 있다는 점을 의미한다. 이를 통해 우리는 원하는 목소리로 더욱 풍부한 표현이 가능한 다양한 TTS 시스템을 쉽게 활용할 수 있게 될 것으로 기대된다.