학사학위 청구논문

추모 서비스를 위한 AI 기반 음성 생성 자동화 시스템에 관한 연구

(Research on AI-Based Automated Voice Generation Systems for Memorial Services)

2023년 12월 17일

숭실대학교 IT대학

전자정보공학부 IT융합전공

백 승 우

학사학위 청구논문

추모 서비스를 위한 AI 기반 음성 생성 자동화 시스템에 관한 연구

Research on AI-Based Automated
Voice Generation Systems for Memorial Services

지도교수 : 김 동 성

이 논문을 학사학위 논문으로 제출함

2023년 12월 17일

숭실대학교 IT대학
전자정보공학부 IT융합전공
백 승 우

백승우의 학사학위 논문을 인준함

심사위원장 신 오 순 (인)

심 사 위 원 김 동 성 (인)

2023년 12월 17일

숭실대학교 IT대학

감사의 글

논문이 완성되도록 도와주신 모든 분께 감사드립니다.

목 차

표 및 그림 목차(크기 16)

추모 서비스를 위한 AI 기반 음성 생성 자동화 시스템에 관한 연구

전자정보공학부 IT융합전공 백 승 우
지도교수 김 동 성

본 논문은 추모 서비스를 위한 AI 기반 음성 생성 및 합성 자동화 시스템의 개발에 관한 연구입니다. 첫 번째로, AI 음성 생성 모델인 Tacotron2와 FastSpeech2를 탐구하며, 이들의 원리와 응용에 대해 분석합니다. 두 번째로, AI 음성 합성 모델에서는 Transfer-Learning, Diff-SVC, RVC와 같은 기술들을 다루며, 이들의 기초 및 적용 사례를 설명합니다.

이어서, 자동화 시스템의 설계 및 구현에 대해 논의합니다. 여기에는 FastAPI 기반 API 설계와 CLI 기반 자동화 시스템 설계가 포함되며, 각각의 장점과 응용 시나리오를 개발합니다. 마지막으로, AI 음성 생성 및 합성 자동화 시스템의 구현 과정을 상세히 설명합니다. 이는 AI 음성 생성 프로세스, AI 음성 합성 방법론, CLI 기반 자동화 시스템 구현, 그리고 FastAPI 기반 API 구현 및 테스트를 포함합니다.

이 연구는 추모 서비스에 AI 음성 기술을 적용함으로써, 인간의 감정과 기억을 존중하고 보존하는 새로운 방식을 제시합니다. 기술적 혁신을 통해 추모 문화에 혁신을 가져오고, AI 음성 기술이 인간의 삶에 미치는 영향을 탐구하는 데 기여합니다.

I. 서 론

본 연구는 추모 서비스를 위한 AI 기반 음성 생성 및 합성 자동화 시스템의 개발과 구현에 중점을 두고 있습니다. 디지털 시대의 진화와 함께, 추모 방식 또한 기술적인 발전을 통해 변모하고 있습니다. 이러한 변화의 일환으로, 본 연구는 AI 음성 기술의 발전을 활용하여, 사랑하는 사람들의 기억을 보다 생동감 있고 개인화된 방식으로 보존하는 새로운 접근법을 탐색합니다.

연구의 2장은 AI 음성 생성 기술에 관한 것으로, Tacotron2와 FastSpeech 모델을 중심으로 그 기초 원리, 작동 원리 및 장점을 분석합니다. 이러한 모델들은 음성 생성 과정의 핵심 요소로서, 높은 질의 음성 출력을 가능하게 하는 기술적 기반을 제공합니다.

3장에서는 AI 음성 합성 기술에 초점을 맞춥니다. 여기에는 Transfer-Learning, Diff-SVC, RVC와 같은 최신 기술들이 포함되며, 각각의 기술에 대한 기초적인 이해, 구현 방법 및 장점을 다룹니다. 이러한 기술들은 음성의 다양성과 개성을 풍부하게 하여, 추모 서비스에 더 깊은 감동과 의미를 더할 수 있습니다.

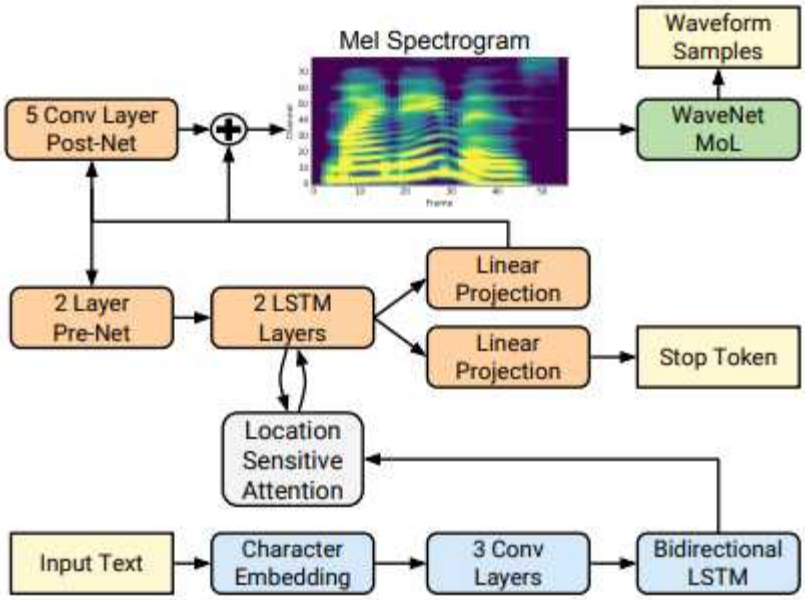
4장은 자동화 시스템의 설계와 시나리오에 관한 것입니다. FastAPI 기반의 API 설계와 CLI 기반의 자동화 시스템 설계 방법을 소개하고, 실제 응용 시나리오의 개발 과정을 설명합니다. 이러한 시스템 설계는 추모 서비스를 위한 AI 음성 생성 및 합성 프로세스의 효율성과 접근성을 크게 향상시킵니다.

5장은 AI 음성 생성 및 합성 자동화 시스템의 실제 구현 과정을 설명합니다. 이는 AI 음성 생성 프로세스와 합성 방법론, 그리고 시스템의 구현 및 테스트를 포함합니다. 이 과정에서 개발된 시스템은 추모 서비스를 위한 새로운 경험을 제공하며, 사용자에게 더욱 개인화되고 의미 있는 추모 방식을 가능하게 합니다.

이 연구는 기술적 혁신을 통해 인간의 감정과 기억을 존중하고 보존하는 새로운 방법을 모색하며, 이를 통해 추모 문화에 새로운 지평을 열고자 합니다. 또한, 이 연구는 미래의 연구 방향을 제시하며, AI 음성 기술이 인간의 삶에 더 깊은 영향을 미칠 수 있는 가능성을 탐구합니다.

II. AI 음성 생성 기술

II-1. Tacotron2



개요

Tacotron2는 Google에서 개발한 최신 음성 합성(Text-to-Speech, TTS) 시스템으로, 자연스러운 인간의 목소리를 모방하는 데 초점을 맞추고 있습니다. 이 모델은 딥 러닝 기술을 기반으로 하며, 텍스트 입력을 받아 고품질의 음성 출력을 생성합니다. Tacotron2의 핵심 목표는 자연스러운 발음, 강세, 그리고 인간과 유사한 억양을 생성하는 것입니다.

작동 원리

Tacotron2는 두 가지 주요 구성 요소로 이루어져 있습니다: 시퀀스-투-시퀀스(sequence-to-sequence) 모델과 WaveNet 기반 보코더입니다. 시퀀스-투-시퀀스 모델은 텍스트 입력을 멜 스펙트로그램(mel-spectrogram)으로 변환하고, 이 멜 스펙트로그램은 WaveNet 보코더에 의해 원시 오디오 신호로 변환됩니다.

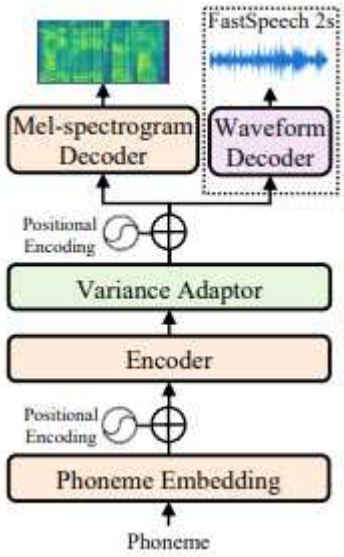
- 시퀀스-투-시퀀스 모델: 텍스트를 멜 스펙트로그램으로 변환하는 과정에서, 모델은 텍스트의 음절과 각 음절의 발음 방식을 학습합니다. 이는 텍스트의 발음과 리듬을 이해하고, 이를 오디오 신호에 매핑하는 데 중요합니다.

- WaveNet 보코더: 멜 스펙트로그램을 원시 오디오 신호로 변환하는 역할을 합니다. WaveNet은 딥 러닝을 기반으로 한 오디오 생성 모델로, 매우 사실적인 인간의 목소리를 생성할 수 있습니다.

장점

Tacotron2의 가장 큰 장점은 그 자연스러움과 유연성입니다. 이 시스템은 다양한 문맥과 억양, 강세를 자연스럽게 표현할 수 있으며, 인간의 목소리와 매우 유사한 오디오를 생성합니다. 또한, 다양한 언어와 방언에 대한 학습이 가능하며, 특정 목소리의 특성을 모방하는 데에도 효과적입니다.

II-2. FastSpeech2



개요

FastSpeech2는 음성 합성 분야에서 주목받는 최신 기술 중 하나로, 텍스트를 음성으로 변환하는 과정을 효율적으로 수행합니다. 이 모델은 기존의 Tacotron2 같은 모델들보다 빠른 속도와 높은 안정성을 제공하며, 자연스러운 음성 품질을 유지합니다. FastSpeech2의 목표는 더 빠른 처리 속도와 더 낮은 오류율로 고품질의 음성 합성을 달성하는 것입니다.

작동 원리

FastSpeech2는 비자기 회귀(non-autoregressive) 변환 모델을 기반으로 하며, 이는 각 시간 단계에서 출력을 병렬로 생성할 수 있어 처리 속도를 대폭 향상시킵니다. 또한, 모델은 멜 스펙트로그램을 직접 생성하고, 이를 보코더로 변환하여 최종 음성 신호를 생성합니다.

- 비자기 회귀 변환: FastSpeech2는 음절과 음절 간의 관계를 동시에 고려하여 멜 스펙트로그램을 생성합니다. 이 접근 방식은 연산을 병렬로 수행할 수 있게 하여 음성 생성 속도를 증가시킵니다.

- 멜 스펙트로그램 생성: 생성된 멜 스펙트로그램은 음성의 품질과 자연스러움을 결정하는 중요한 요소입니다. FastSpeech2는 높은 정확도의 멜 스펙트로그램을 생성하여 더 자연스러운 음성을 달성합니다.

장점

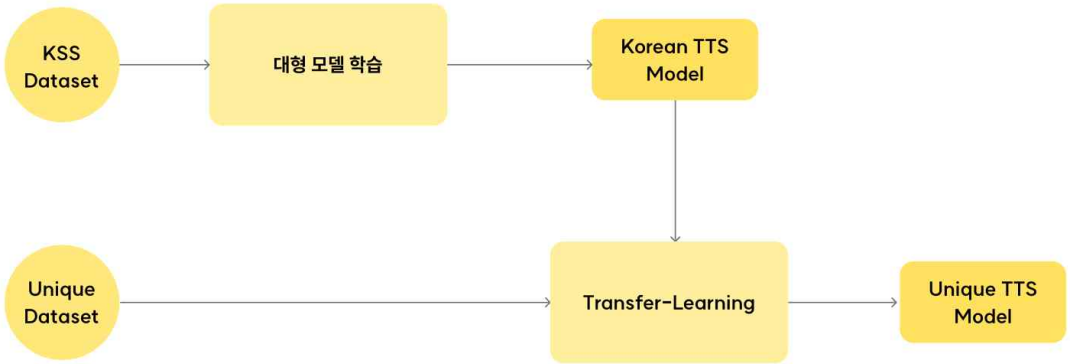
FastSpeech2의 주요 장점은 빠른 처리 속도와 높은 안정성입니다. 이는 특히 대규모 음성 합성 작업에서 중요하며, 실시간 음성 변환 시스템에도 적합합니다. 또한, FastSpeech2는 다양한 억양과 발음을 정확하게 모델링할 수 있어, 보다 자연스러운 음성 합성이 가능합니다.

결론

Tacotron2와 Fastspeech2 두 음성 합성 시스템을 비교한 결과, 성능 측면에서 Tacotron2가 우수하다는 결론에 도달했습니다. Tacotron2는 음성 합성의 자연스러움과 명료성에서 뛰어난 성능을 보였으며, 특히 인간의 음성에 가까운 발음과 감정 표현력이 높게 평가되었습니다. 반면, Fastspeech2 역시 빠른 처리 속도와 효율성에서 장점을 가지고 있지만, 최종적인 음성의 자연스러움과 정확성 면에서 Tacotron2에 미치지 못하는 것으로 나타났습니다. 따라서 성능의 우수성을 최우선 기준으로 삼을 때 Tacotron2가 더 적합한 선택이라고 할 수 있습니다.

III. AI 음성 합성 기술

III-1. Transfer-Learning



개요

전이 학습(Transfer-Learning)은 인공지능 및 머신 러닝 분야에서 널리 사용되는 기술로, 한 영역에서 학습한 지식을 다른 영역의 문제 해결에 적용하는 방법입니다. 이 방법은 특히 음성 합성과 같이 대규모 데이터가 필요한 분야에서 유용하며, 제한된 데이터로도 효과적인 모델을 학습시킬 수 있도록 도와줍니다.

작동 원리

전이 학습의 핵심은 이미 학습된 모델(일반적으로 큰 데이터셋에서 학습된 모델)을 취해, 이를 새로운, 종종 더 작은 데이터셋에 적용하는 것입니다. 이 과정에서, 기존 모델의 지식을 유지하면서 새로운 데이터에 특화된 학습을 추가로 수행합니다. 예를 들어, 일반적인 음성 데이터로 훈련된 모델을 특정 인물의 목소리 스타일에 맞게 조정하는 것이 가능합니다.

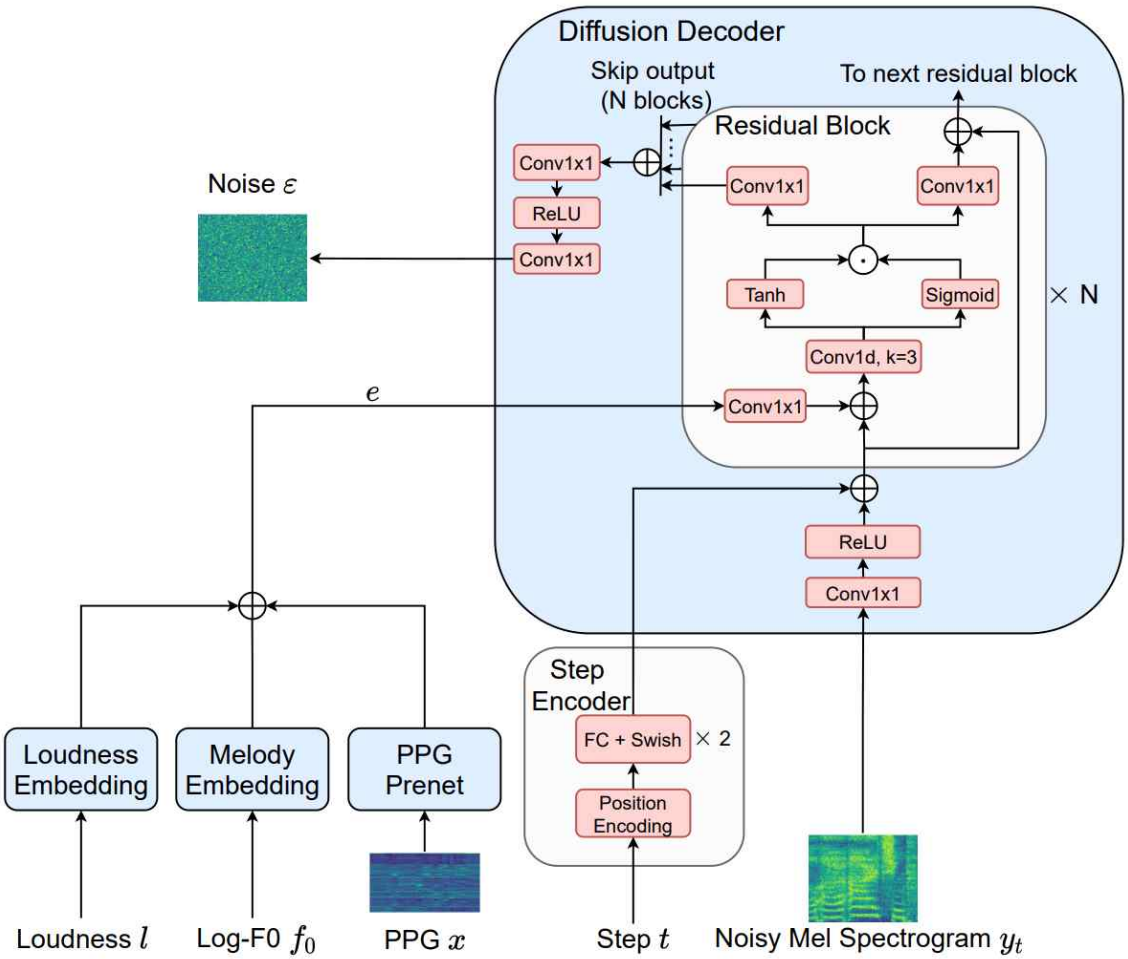
- 기존 모델 재사용: 대규모 데이터셋에서 미리 학습된 모델을 사용함으로써, 학습 시간을 단축시키고 높은 성능을 유지할 수 있습니다.

- 미세 조정(Fine-tuning): 새로운 데이터셋에 대해 모델의 일부를 미세 조정하여 특정 목적에 맞게 최적화합니다. 이는 적은 양의 데이터로도 효과적인 학습을 가능하게 합니다.

장점

전이 학습은 학습 시간 및 데이터 요구량을 크게 줄여줍니다. 이는 특히 데이터가 제한적인 경우나, 빠른 모델 개발이 필요한 상황에서 유리합니다. 또한, 전이 학습은 모델의 일반화 능력을 향상시켜, 새로운 상황에 대한 적응력을 강화합니다

III-2. Diff-SVC



개요

Diff-SVC (Diffusion Probabilistic Model for Singing Voice Conversion)는 노래의 목소리를 다른 가수의 목소리로 변환하는 최신 기술입니다. 이 기술은 노래의 내용(content)과 멜로디를 그대로 유지하면서, 목소리의 특성만을 변경합니다. 이는 최근 SVC (Singing Voice Conversion) 분야에서 주목받는 혁신적인 접근법입니다.

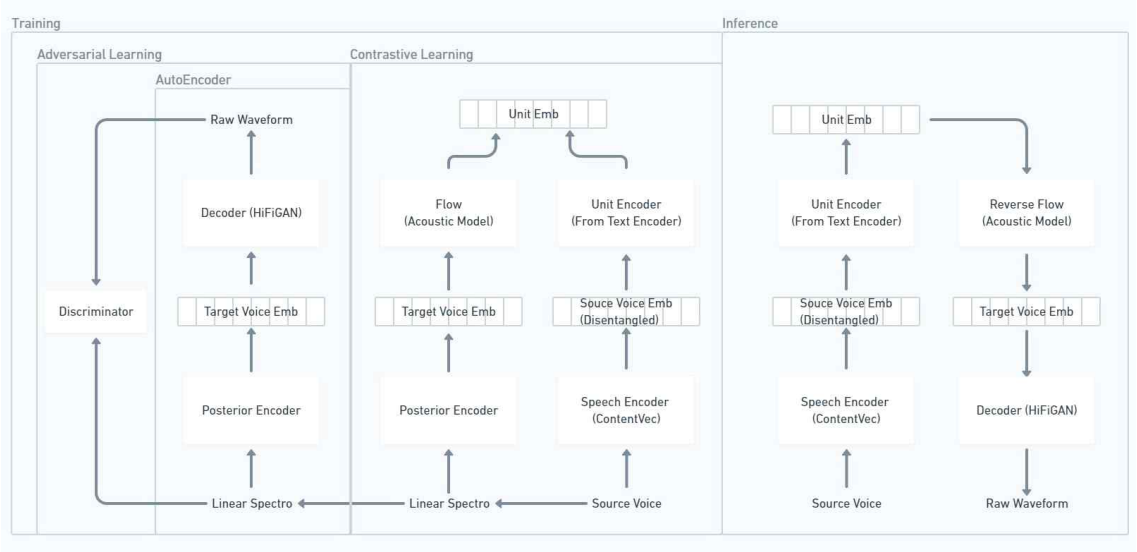
작동 원리

Diff-SVC는 두 가지 주요 구성 요소로 작동합니다: ASR (Automatic Speech Recognition) 모델과 diffusion model입니다. ASR 모델은 노래에서 콘텐츠 feature를 추출하며, 이는 end-to-end 모델이거나 hybrid HMM-DNN 모델일 수 있습니다. 추출된 feature는 diffusion model을 통해 mel spectrogram과 같은 spectral feature로 변환되며, 이후 waveform을 생성합니다. Diffusion model의 고유한 특징은 데이터에 점진적으로 Gaussian noise를 더하는 forward process와 이를 제거하여 원래 데이터를 복구하는 reverse process로 구성됩니다.

장점

Diff-SVC의 가장 큰 장점은 목소리 변환의 자연스러움과 유사성입니다. 기존의 SVC 시스템과 비교하여, Diff-SVC는 더욱 개선된 음질과 변환 정확도를 제공합니다. 이 기술은 음성의 다양한 특성을 효과적으로 모델링하고, 자연스러운 음색 변환을 가능하게 합니다.

Ⅲ-3. RVC



개요

RVC, 즉 Retrieval-based Voice Conversion은 고급 알고리즘과 딥러닝 기법을 결합한 AI 모델로, 실시간 인터랙티브하고 다이내믹한 목소리 복제를 생성합니다. 이 기술은 특히 음성 변환 분야에서 혁신적인 접근법을 도입하고 있으며, 자연스러운 인간 음성의 반응을 기반으로 생성됩니다.

작동 원리

RVC는 자연어 처리(NLP)와 음성 인식 기술을 통합하여, 실시간으로 상호작용하고 다이나믹한 음성 클론을 생성합니다. 이 모델은 더 나은 목소리 클로닝 정확도와 품질을 위해 개선된 알고리즘을 사용하는 RVC v2로 발전되었습니다.

장점

RVC의 주요 장점은 높은 품질의 음성 변환과 실시간 변환 기능입니다. 전통적인 음성 변환기와 달리, RVC는 단순히 음성의 피치나 속도를 변경하는 것이 아니라, 더 자연스럽고 인간 같은 결과를 제공합니다. 이 기술은 다양한 목소리를 선택할 수 있는 광범위한 라이브러리를 제공하여 사용자가 원하는 목소리로 변환할 수 있습니다.

결론 - RVC 선택

	Transfer-Learning	Diff-SVC	RVC
기술	기존 모델에 전이 학습	Stable-Diffusion	음성 변조
정확도	높음	보통	보통
학습시간	매우 느림	빠름	매우 빠름
필요 데이터셋	고음질 많은 데이터	고음질 적당한 데이터	저음질 가능 적은 데이터

본 연구에서 Transfer-Learning, Diff-SVC, 그리고 RVC 중에서 속도 측면을 고려하여 RVC를 선택했습니다. 이 결정의 주된 이유는 다음과 같습니다.

- 속도와 효율성: RVC (Retrieval-based Voice Conversion)는 실시간 음성 변환에 탁월한 성능을 보입니다. 이는 RVC가 강력한 음성 인식과 자연어 처리 기술을 통합하여 빠른 처리 속도를 제공하기 때문입니다. 비교적 적은 계산 자원으로도 높은 품질의 음성 변환을 가능하게 하는 RVC의 효율성은 본 연구의 목적에 부합합니다.

- 사용자 경험 향상: RVC 기술은 사용자에게 더 나은 경험을 제공합니다. 실시간 음성 변환 기능은 사용자가 원하는 음성 스타일로 즉각적인 변환을 경험할 수 있게 해줍니다. 이는 특히 추모 서비스와 같이 감정적 연결이 중요한 분야에서 사용자 만족도를 높이는 데 기여합니다.

- 응용 프로그램의 다양성: RVC의 빠른 처리 속도는 다양한 응용 프로그램에서의 활용 가능성을 높입니다. 예를 들어, 추모 서비스 외에도 음악 제작, 오디오북 제작, 가상 보조원 등 다양한 분야에서 RVC의 활용이 가능합니다.

이러한 이유로, RVC는 본 연구에서 추구하는 AI 기반 음성 생성 자동화 시스템의 속도와 효율성 측면에서 가장 적합한 선택으로 판단되었습니다.

Ⅳ. 자동화 시스템의 설계와 시나리오

Ⅳ-1. CLI 기반 자동화 시스템 설계

개요

CLI (Command Line Interface) 기반 자동화 시스템 설계는 RVC (Retrieval-based Voice Conversion) 기술을 활용하여 목소리 변환 프로세스를 자동화하는 방법입니다. 이 시스템은 사용자가 명령줄 인터페이스를 통해 음성 변환 작업을 효율적으로 관리하고 실행할 수 있게 합니다.

설계의 핵심

CLI 기반 자동화 시스템은 사용자가 명령어를 입력하여 음성 변환 작업을 제어할 수 있도록 설계되었습니다. 이 시스템은 다음과 같은 주요 요소로 구성됩니다.

- 음성 변환 명령어: 사용자는 CLI를 통해 RVC 기술을 사용하여 음성 변환을 시작하고 제어할 수 있는 명령어를 입력합니다.

- 파라미터 조정: 음성 변환의 특성 및 파라미터(예: 음성의 피치, 속도, 스타일 등)는 사용자의 요구에 맞게 조정될 수 있습니다.

- 음성 파일 관리: 변환된 음성 파일은 자동으로 관리되며, 사용자는 필요에 따라 이 파일을 저장하거나 다른 애플리케이션으로 전송할 수 있습니다.

- 효율적인 프로세스 관리: 자동화 시스템은 음성 변환 프로세스를 최적화하여 빠르고 효율적인 변환을 가능하게 합니다.

시스템의 장점

CLI 기반 자동화 시스템은 다음과 같은 장점을 가집니다.

- 사용자 친화적: 명령줄 인터페이스는 사용자가 쉽게 음성 변환 작업을 제어할 수 있도록 합니다.

- 유연성: 다양한 사용자 요구에 맞춰 음성 변환 파라미터를 조정할 수 있는 유연성을 제공합니다.

- 효율성: 자동화 시스템은 음성 변환 프로세스를 최적화하여 빠른 변환 속도와 높은 품질을 보장합니다.

Ⅳ-2. FastAPI 기반 API 설계

개요

FastAPI 기반 API 설계는 현대적인, 빠르고, 사용하기 쉬운 API를 제공하는 데 초점을 맞춘 개발 접근법입니다. FastAPI는 Python으로 작성된 웹 프레임워크로, 높은 성능, 쉬운 코드 작성, 그리고 자동화된 문서 생성 기능을 제공합니다.

설계의 핵심

FastAPI 기반 API 설계의 주요 목적은 효율적인 데이터 교환과 고성능 웹 서비스를 제공하는 것입니다. 이 설계는 다음과 같은 주요 요소를 포함합니다:

- 경량화 및 빠른 성능: FastAPI는 비동기 프로그래밍을 지원하여 빠른 응답 시간과 높은 동시성 처리 능력을 제공합니다.

- 자동 문서화: FastAPI는 OpenAPI와 Swagger UI를 사용하여 API의 모든 끝점(endpoints)에 대한 문서를 자동으로 생성합니다.

- 타입 힌트와 데이터 검증: Python의 타입 힌트를 사용하여 데이터의 유효성 검증과 API 설계의 간소화를 도모합니다.

- 쉬운 오류 처리: 사용자 정의 예외 처리를 통해 API 사용 중 발생할 수 있는 오류를 용이하게 관리할 수 있습니다.

장점

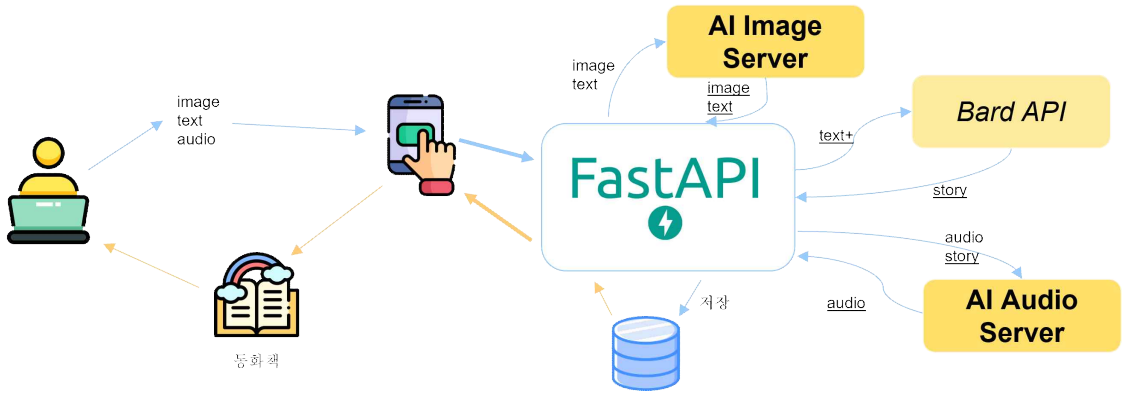
FastAPI 기반의 API 설계는 다음과 같은 장점을 가집니다.

- 고성능: FastAPI는 Starlette과 Pydantic을 기반으로 하여 높은 성능을 제공합니다.

- 개발 용이성: 직관적인 API 설계와 자동화된 문서화로 개발 과정이 간소화됩니다.

- 강력한 데이터 처리: 타입 힌트를 통한 데이터 검증과 직렬화를 통해 안정적인 데이터 처리가 가능합니다.

Ⅳ-3. 개발 시나리오



개요

개발 시나리오는 FastAPI를 사용한 웹 서비스와 RVC 기술을 활용한 음성 변환 기능의 통합 과정을 설명합니다. 이 시나리오는 추모 서비스용 음성 변환 시스템의 개발 및 배포를 목표로 합니다.

시나리오 설계

1. 시스템 구축의 시작: FastAPI를 사용하여 추모 서비스를 위한 웹 서비스의 기본 구조를 구축합니다. 이 단계에서는 사용자가 쉽게 접근할 수 있는 인터페이스와 효율적인 백엔드 로직을 설계합니다.

2. RVC 통합: 추모 서비스의 핵심 기능으로, RVC 음성 변환 기술을 웹 서비스에 통합합니다. 사용자는 사랑하는 사람의 목소리 특성을 선택하거나 제공할 수 있습니다.

3. 음성 변환 기능 구현: 사용자가 업로드한 음성 샘플을 기반으로, 사랑하는 사람의 목소리 스타일로 변환합니다. 이는 감정적인 연결과 추모 경험을 강화합니다.

4. API 엔드포인트 개발: 사용자가 음성 샘플을 쉽게 업로드하고 변환된 결과를 받을 수 있도록 API 엔드포인트를 개발합니다.

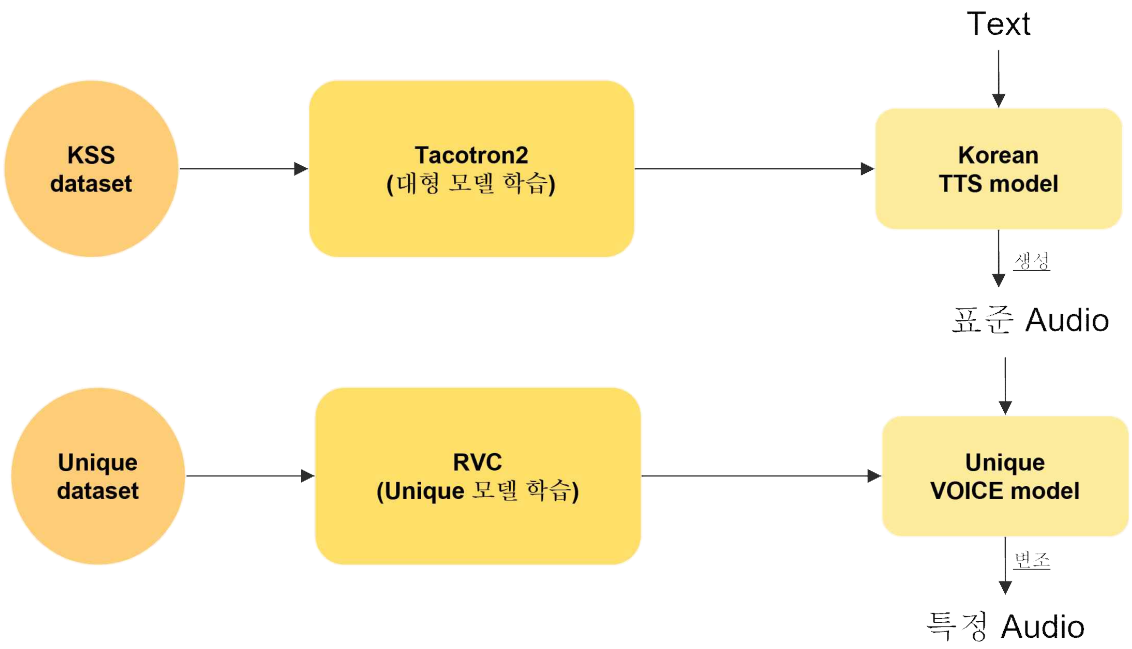
5. 사용자 인터페이스 개발: 사용자가 웹 사이트를 통해 음성 샘플을 업로드하고, 추모 서비스를 이용할 수 있는 친숙한 인터페이스를 제공합니다.

6. 시스템 테스트 및 최적화: 개발된 시스템을 다양한 시나리오에서 테스트하여 사용자 경험을 개선하고 시스템 성능을 최적화합니다.

7. 배포 및 지속적 관리: 시스템이 안정적으로 작동하는 것을 확인한 후, 서비스를 배포하고 지속적인 관리를 수행합니다.

Ⅴ. AI 음성 생성 자동화 시스템 구현

Ⅴ-1. AI 음성 생성 프로세스



AI 음성 생성의 전반적 흐름

1. 대형 음성 모델 생성: AI 음성 생성의 첫 단계는 대형 음성 모델을 구축하는 것입니다. 이 과정에서는 END-to-END 방식의 대표 모델인 Tacotron2를 사용합니다. Tacotron2는 높은 정확도를 제공하지만 긴 학습 시간이 필요합니다.

2. 프리 트레인드 모델 활용: Tacotron2의 긴 학습 시간 문제를 해결하기 위해 추가 모델 생성을 중단하고, 사전 훈련된(프리 트레인드) 모델을 활용하기로 결정했습니다. 이 접근 방식은 학습 시간을 단축시키고 프로젝트의 효율성을 높입니다.

3. 특정 음성 모델 생성: 다음 단계는 특정 음성 모델을 생성하는 것입니다. 이를 위해 전이 학습, Diff-SVC, 그리고 RVC 중에서 RVC를 선택했습니다. RVC는 다른 모델에 비해 정확도는 평균 수준이지만, 저음질의 적은 데이터로도 훈련이 가능하며 빠른 처리 속도를 제공합니다.

AI 음성 생성의 상세 과정

1. 텍스트 입력과 표준 음성 파일 생성: 사용자로부터 입력된 텍스트는 사전 훈련된 대형 음성 모델(Tacotron2)을 통해 표준 음성 파일로 변환됩니다. 이 과정에서 모델은 텍스트의 의미와 감정을 파악하여 자연스러운 음성으로 출력합니다.

2. RVC를 통한 음성 변조: 생성된 표준 음성 파일은 RVC 기술을 사용하여 사용자가 원하는 파형으로 변조됩니다. RVC는 입력된 음성 데이터를 분석하여, 특정한 스타일이나 특성을 갖는 목소리로 변환하는 과정을 거칩니다.

3. AI 음성 최종 출력: 변조된 음성 파일은 최종 AI 음성으로 제공됩니다. 이 음성은 추모 서비스와 같은 응용 프로그램에서 사용될 수 있으며, 감정적 연결을 강화하는 데 기여합니다.

V-2. CLI 기반 자동화 시스템 구현

Q7:How to train and infer without the WebUI?

Training script:
You can run training in WebUI first, and the command-line versions of dataset preprocessing and training will be displayed in the message window.

Inference script:
https://huggingface.co/lj1995/VoiceConversionWebUI/blob/main/myinfer.py

e.g.

```
runtime\python.exe myinfer.py 0 "E:\codes\py39\RVC-beta\todo-songs\1111.wav"
"E:\codes\py39\logs\mi-test\added_IVF677_Flat_nprobe_7.index" harvest "test.wav"
"weights/mi-test.pth" 0.6 cuda:0 True

f0up_key=sys.argv[1]
input_path=sys.argv[2]
index_path=sys.argv[3]
f0method=sys.argv[4]#harvest or pm
opt_path=sys.argv[5]
model_path=sys.argv[6]
index_rate=float(sys.argv[7])
device=sys.argv[8]
is_half=bool(sys.argv[9])
```

개요

CLI (Command Line Interface) 기반 자동화 시스템 구현은 사용자가 명령 줄 인터페이스를 통해 음성 변환 작업을 실행하고 제어할 수 있도록 하는 과정입니다. 이 시스템은 RVC (Retrieval-based Voice Conversion) 기술을 사용하여 특정 음성 파일을 변환하는 데 필요한 명령어와 파라미터를 포함합니다.

구현 단계

1. 데이터셋 전처리 및 트레이닝 스크립트: WebUI에서 훈련을 수행한 후, 명령 줄 버전의 데이터셋 전처리와 트레이닝 명령이 메시지 창에 표시됩니다. 사용자는 이 정보를 기반으로 CLI에서 훈련을 진행할 수 있습니다.

2. 추론 스크립트 사용: 추론을 위한 스크립트는 Hugging Face의 저장소에서 제공됩니다 (예시: myinfer.py). 이 스크립트를 사용하여 명령 줄에서 음성 파일 변환을 실행할 수 있습니다.

3. 명령어 구조: 다음과 같은 명령어 구조를 사용하여 추론을 수행합니다.

CLI Command

runtime\python.exe myinfer.py [f0up_key] "[input_path]" "[index_path]" [f0method] "[opt_path]" "[model_path]" [index_rate] [device] [is_half]

여기서 각 인자는 다음과 같은 의미를 가집니다:

- f0up_key: 피치 조정 키
- input_path: 입력 음성 파일 경로
- index_path: 인덱스 파일 경로
- f0method: 피치 추출 방법 (harvest 또는 pm)
- opt_path: 출력 파일 경로
- model_path: 모델 파일 경로
- index_rate: 인덱스 비율
- device: 사용할 디바이스 (예: cuda:0)
- is_half: half precision 사용 여부

실행 및 결과

위의 명령어를 사용하여 사용자는 특정 음성 모델을 사용해 입력된 음성 파일을 원하는 형태로 변환할 수 있습니다. 이 과정은 실시간이나 배치 처리 형태로 진행될 수 있습니다.

Ⅴ-3. FastAPI 기반 API 구현

개요

본 연구의 웹 서버는 FastAPI 기반으로 구현되며, 사용자로부터 텍스트를 수신하여 AI 음성을 생성한 후, 생성된 음성을 플랫폼 서버로 전송하는 기능을 가집니다. 이 구현은 효율적인 음성 처리와 빠른 네트워크 통신을 목표로 합니다.

구현 절차

1. API 라우팅 설정: FastAPI를 사용하여 음성 생성 요청을 처리하는 API 라우트를 설정합니다. 이 라우트는 사용자로부터 텍스트 데이터를 수신하고, 해당 데이터를 처리하여 음성 생성 요청을 시작합니다.

2. 음성 생성 처리: 수신된 텍스트는 AI 음성 생성 모듈(Tacotron2, RVC 등)에 전달되어 처리됩니다. 이 과정에서 텍스트는 자연스러운 음성으로 변환됩니다.

3. 비동기 처리: 음성 생성 작업은 비동기적으로 수행되어, 서버의 응답 속도를 최적화합니다. FastAPI의 비동기 기능을 활용하여 서버 부하를 최소화하고, 동시 요청 처리 능력을 향상시킵니다.

4. 결과 전송: 음성 생성이 완료되면, 생성된 음성 파일은 플랫폼 서버로 전송됩니다. 이는 FastAPI의 HTTP 클라이언트를 통해 안정적으로 수행됩니다.

5. 오류 처리 및 로깅: 모든 과정에서 발생할 수 있는 오류는 적절히 처리되며, 서버의 작업 로그는 시스템의 투명성과 유지보수를 위해 기록됩니다.

6. API 문서화: FastAPI의 자동 문서화 기능을 활용하여, API의 사용 방법과 파라미터를 명확하게 문서화합니다. 이는 개발자와 최종 사용자가 API를 쉽게 이해하고 사용할 수 있도록 돕습니다.

Ⅵ. 결 론

본 연구에서는 추모 서비스를 위한 AI 기반 음성 생성 및 합성 자동화 시스템의 개발을 다룹니다. 이 시스템은 인간의 감정과 기억을 보존하고 추모하는 새로운 방식을 제시합니다. 연구 과정에서는 성능 측면에서 우수한 Tacotron2와 처리 속도가 빠른 RVC를 선택했습니다. Tacotron2는 높은 음질과 자연스러운 음성 생성 능력을 제공하는 반면, RVC는 빠른 음성 변환 처리 속도를 가지고 있어, 두 기술의 결합이 시스템의 효율성과 효과성을 극대화합니다. 이 연구는 FastAPI 기반의 API 설계와 CLI 기반 자동화 시스템을 통해 음성 생성 및 합성 프로세스를 개선하고, 사용자의 접근성을 높였습니다. 이를 통해 추모 문화에 혁신을 가져오고 AI 음성 기술이 인간의 삶에 미치는 영향을 탐구하는 중요한 기여를 합니다.

참고문헌

[1] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the Interspeech 2017.

[2] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Agiomyrgiannakis, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[3] Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & Shoeybi, M. (2017). Deep Voice: Real-time Neural Text-to-Speech. arXiv preprint arXiv:1702.07825.

[4] Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. arXiv preprint arXiv:1609.03499.

[5] Ren, Y., Ruan, X., Tan, J., Qin, T., Zhao, Z., Liu, T., & Gao, S. (2020). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. arXiv preprint arXiv:2006.04558.

[6] Liu, T., Ren, Y., Zhao, Z., Qin, T., & Gao, S. (2019). FastSpeech: Fast, Robust and Controllable Text to Speech. In Advances in Neural Information Processing Systems (NeurIPS).

[7] Chen, Y., Ling, Z. H., & Liu, L. (2019). End-to-End Text-to-Speech with Neural Attention. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[8] Tan, J., Qin, T., Ren, Y., Zhao, Z., Liu, T., & Gao, S. (2021). Survey on Deep Learning for Text-to-Speech Synthesis. arXiv preprint arXiv:2106.15561.

[9] Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

[10] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A Survey of Transfer Learning. Journal of Big Data, 3(1), 9.

[11] Torrey, L., & Shavlik, J. (2010). Transfer Learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques (pp. 242-264). IGI Global.

[12] Ruder, S. (2019). Neural Transfer Learning for Natural Language Processing. PhD Thesis, National University of Ireland, Galway.

[13] Liu, S., Ren, Y., Yu, C., Chen, Z., & Zhou, K. (2021). DiffSVC: A Diffusion Probabilistic Model for Singing Voice Conversion. arXiv preprint arXiv:2105.13871.

[14] RVC-Project. (2023). Retrieval-based-Voice-Conversion-WebUI (Version 1006v2). GitHub.
https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI