TTS DATA

음성 합성을 위한 데이터 준비

1. 음성 합성

1.1. 음성 합성이란 무엇인가?

인위적으로 사람의 소리를 합성하여 문자(TEXT)를 음성(SPEECH)으로 변환해 주는 시스템

• 알고리즘의 대략적인 구조

2. 데이터의 중요성

2.1. 쉽게 이해하는 음성 합성 구조

알고리즘은 고정적인 일정한 성능을 제공하지만, 아웃풋의 퀄리티의 결정적인 차이는 인풋 데이터에 의함
파라미터 등은 조정할 것이 적어 데이터보다는 영향을 덜 미치지만 영향을 미칠 때는 크게 미침

- 에어프라이어 = 알고리즘
- 식재료 = 데이터
- 온도 / 시간 조절 = Parameter 등 조절

2. 데이터의 중요성

2.2. 데이터에 따른 아웃풋의 퀄리티 차이

- 낮은 퀄리티의 결과물

"존경하는 국민 여러분, 대한민국 대통령 문재인 입니다.
요즘 딥페이크 기술을 악용해서 만든 가짜 뉴스가 사회적으로 문제가 되고 있습니다.
대한민국 정부는 이러한 문제를 해결하기위해 인공지능 기술을 적극적으로 활용하고,
기술 개발 회사에 대한 지원을 확대해나갈 것입니다."

- 자연스러운 발화 합성이 가능한 높은 퀄리티의 결과물

독일 함부르크 공항에 막 착륙한 비행기 안에서, 잔잔히 흐르는 비틀스의 노르웨이의 숲을 들으며, 남자는 오랜 세월을 거슬러 올라, 간절한 부탁과 그 부탁을 남긴 한 여자를 추억합니다.

2. 데이터의 중요성

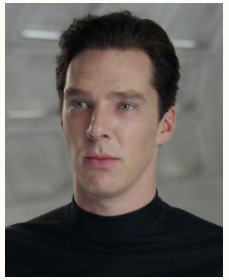
2.2. 데이터에 따른 아웃풋의 퀄리티 차이

- 자연스러운 발화 합성이 가능한 높은 퀄리티의 결과물

느그들 보고 싶어 멧자 적는다., 추위에 별 일 없드나? , 내사 방 따시고, 밥잘 묵으이 걱정 없다., 건너말 작은 할배 제사가 멀지 않았다,. 잊아뿌지 마라,. 몸들 성커라. 돈 멧 닢 보낸다. 공책사라.

배우 이제훈입니다. 제가 출연한 사냥의 시간이 얼마 전, 넷플릭스에 공개되었습니다. 많은 시청 부탁드립니다. 감사합니다.

제가 이천팔년 러시아 상트대학 연설에서 이런 말을 했었습니다. 미래는 새로운 꿈을 갖고, 불가능에 도전하는 자들의 것입니다. 이 말을 가슴에 새기고 저도 미래를 향해 나아가겠습니다.



My soul was struggling. Caterina ended the struggle.

2. 데이터의 중요성

2.3. 데이터를 통한 문제의 해결

- **데이터로 인해 발생 가능한 문제**

  1) 특정 character에 따른 발화 / 발음 문제
  2) 노이즈로 인한 음질 저하
  3) 데이터 편중으로 인한 발화 불가

  ⋮

- **데이터를 활용한 문제 해결의 과정**

문제점 확인	데이터 구조 확인	스크립트 확인	커스터마이징
데이터가 갖는 문제점 확인	텍스트 스크립트 구조 확인	음성 및 텍스트 데이터의 매칭 정확성 확인	데이터를 문제점 해결이 가능하도록 커스터마이징

2. 데이터의 중요성

2.3. 데이터를 통한 문제의 해결

• 데이터를 활용한 문제 해결 예시 - 재난 모델

1

• 문제점

1) 문단의 마무리 마침표 여부
- 문단의 마지막에 마침표가 없을 때 모델이 터짐
- 마침표 생성시 정상 발화)

2) 붙여 읽기
- 예) 전북북부앞바다중 → 모델 터짐
  전북 북부 앞바다 중 → 정상

• 해결방법

1) 인위적 마침표 제거
- 마침표 붙은 문장 = 전체의 81%
- 아무것도 안 붙은 데이터 증량

2) 텍스트 데이터 강제 붙임
- 예) 아치판넬 구조 지붕에
  → 아치판넬구조지붕에

2

• 문제점

1) 단어 무시
- 문장 내에 있는 어절을 무시

2) 과도한 끊어 읽기
- 2~3 어절 단위로 끊어 읽음

• 해결방법

1) 오탈자 확인
- 스크립트 매칭 확인

2) 쉼표 및 마침표 수정
- 쉼표와 마침표가 일관성을 유지하도록 수정

3

• 문제점

1) 마침표 여부에 따른 발음 오류
- 예) 소설. [소선] / 소설, [소설]

2) 특정 단어 발음 오류
- 예) ~것 [것 가]
  ~하지마 [하지만]

• 해결방법

인위적 데이터 증량을 통해 나타나는 두 가지의 문제점을 동시 해결

3. 데이터 구축

3.1. 음성 데이터 수집

- **높은 퀄리티의 음성 합성을 위해 음성 데이터 선정 전 고려해야하는 사항**

1) 노이즈 여부
   - 노이즈가 섞여있을 경우, 음성 합성 시 노이즈가 섞여 나옴
2) 일관성 있는 인토네이션
   - 일관성을 갖지 않는 다양한 인토네이션이 섞여있을 경우 오히려 퀄리티가 저하되기도 함

- **음성 데이터 수집 방법**

1) 직접 녹음
   - 소음이 최대한 발생하지 않는 환경에서의 녹음 필수 / 최소 400 ~ 500 문장 이상 필요
2) 오픈 데이터 활용
   - 데이터 확인 후 정제 필요
3) 데이터 크롤링
   - 유튜브, 오디오북, 팟캐스트, 라디오, 티비 등 다양한 스펙트럼의 데이터 수집 가능
   - 데이터에 대한 퀄리티 보장 불가 및 데이터 정제 시 장시간 소요

3. 데이터 구축

3.2. 음성 데이터 정제 방법

- **일반적인 정제법**

1) 데이터 길이
   - 일반적으로 2 ~ 10초 내외 (1~2문장)
2) 노이즈 정제
   - 음성에 섞인 노이즈 최대한 정제
3) 발음 확인
   - 부정확하게 발화되는 발음 정제 필요
   - 데이터를 다 버리지 말고 불필요한 부분만 정제
4) 음성 파일 인코딩
   - wav, 22k (22050)

- **더 나은 output을 위해 정성을 기울이면 좋은 부분**

1) 원하는 발화 형태가 있을 경우, 일정한 인토네이션 유지 필수
   - 의문문에서 발화의 끝이 올라가는 형태가 필요한 경우, 의문문에 매칭된 음성 데이터의 끝이 올라간 형태일 때 발화 안정성 ↑
2) 쉼표 및 마침표의 일정한 기준
   - 쉼표와 마침표를 찍을 때, 일정한 기준을 가지는 것이 좋음
3) 일정한 음성 데이터 볼륨
   - 음성 데이터의 볼륨이 일정하지 않을 경우, 음성 합성이 제대로 되지 않음
   - 필요시 normalizing 작업 진행

3. 데이터 구축

3.3. 텍스트 데이터 정제 방법

- **일반적인 주의점**

1) 음성과 텍스트 매칭 확인
   - 음성에 맞게 텍스트가 매칭되는지 확인 / 개수 매칭 확인
   - 음성이 표준어 혹은 문법에 맞지 않더라도, 음성에 우선하여 텍스트 전사 필수
     예) 음성 [안냥하시요] → 텍스트 [안냥하시요]

2) 한글만 사용
   - 숫자, 영어, 특수문자 등 사용 금지 (사용 시 학습 불가)
     예) 1학년 → 일학년 / newyork에 살아 → 뉴욕에 살아 / 1972년에 → 천구백칠십이년에
   - 특수문자의 경우 [. , ? ! ~ : ‘ “] 사용 가능

3) 스크립트 구성
   - 자연스러운 발화를 위해 최대한 다양하게 스크립트 구성 필요
   - 스크립트가 편향적일 경우, 특정 스크립트 외엔 음성 합성 시 문제 발생 가능
   - 예) 마침표로 끝나는 데이터가 대부분일 경우 → 문장 끝에 다른 특수문자가 오는 경우, 발화 안정성 감소

3. 데이터 구축

3.3. 텍스트 데이터 정제 방법

- **학습용 스크립트 생성**

  - 기본 포맷: 음성 데이터 경로|본문|speakerID　　(파일명 : speakerID.txt)
    예) disaster_trial/wav/disaster_trial_D00001_A00001.wav|안녕하세요|disaster_trial

  - 텍스트 파일 인코딩: UTF-8 without BOM

  - 해당 포맷에 대한 자세한 사항은 첨부파일 확인

disaster_trial/wav/disaster_trial_D00001_A00001.wav|하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여	disaster_trial/wav/disaster_trial_D00001_A00001.wav|하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열|disaster_trial
disaster_trial/wav/disaster_trial_D00001_A00002.wav|한 켤레씩이나„ 두 명뿐 아니라„ 세	disaster_trial/wav/disaster_trial_D00001_A00002.wav|한 켤레씩이나„ 두 명뿐 아니라„ 세 쪽의„ 네 마리가„ 다섯 권은„
disaster_trial/wav/disaster_trial_D00001_A00003.wav|열한 채라„ 열두 번에„ 열세 묶음이	disaster_trial/wav/disaster_trial_D00001_A00003.wav|열한 채라„ 열두 번에„ 열세 묶음이라„ 열네 가지는„ 열다섯 줄쯤
disaster_trial/wav/disaster_trial_D00001_A00004.wav|영 칼로리, 일 미터, 이 와트, 삼 제	disaster_trial/wav/disaster_trial_D00001_A00004.wav|영 칼로리, 일 미터, 이 와트, 삼 제곱미터, 사 킬로미터, 오 센치,
disaster_trial/wav/disaster_trial_D00001_A00005.wav|구 메가바이트, 십 밀리미터, 십일	disaster_trial/wav/disaster_trial_D00001_A00005.wav|구 메가바이트, 십 밀리미터, 십일 평, 십이 데시벨, 십삼 킬로미터
disaster_trial/wav/disaster_trial_D00001_A00006.wav|열여덟 살, 스물아홉 살, 서른 살, 마	disaster_trial/wav/disaster_trial_D00001_A00006.wav|열여덟 살, 스물아홉 살, 서른 살, 마흔한 살, 쉰 두 살, 예순 세 살
disaster_trial/wav/disaster_trial_D00001_A00007.wav|모든 공부가 서로 녹아든 거야, 안	disaster_trial/wav/disaster_trial_D00001_A00007.wav|모든 공부가 서로 녹아든 거야, 안 그래|disaster_trial
disaster_trial/wav/disaster_trial_D00001_A00008.wav|뭔가 손에 쥐는 게 필요해|disaster	disaster_trial/wav/disaster_trial_D00001_A00008.wav|뭔가 손에 쥐는 게 필요해|disaster_trial
disaster_trial/wav/disaster_trial_D00001_A00009.wav|사람들이 죽음에 씌운 어둠과 검정	disaster_trial/wav/disaster_trial_D00001_A00009.wav|사람들이 죽음에 씌운 어둠과 검정빛은 어쩌자고, 그렇게 된 것일
disaster_trial/wav/disaster_trial_D00001_A00010.wav|이상은 그 소설 안에 무수히 많은	disaster_trial/wav/disaster_trial_D00001_A00010.wav|이상은 그 소설 안에 무수히 많은 사건들 가운데, 비교적 이미지
disaster_trial/wav/disaster_trial_D00001_A00011.wav|이쪽의 뒤편에 서 있던 길산이는, 그	disaster_trial/wav/disaster_trial_D00001_A00011.wav|이쪽의 뒤편에 서 있던 길산이는, 그 울음소리를 듣자, 어쩐지 낯
disaster_trial/wav/disaster_trial_D00001_A00012.wav|그러면 떡볶이에 만두하고, 달걀은	disaster_trial/wav/disaster_trial_D00001_A00012.wav|그러면 떡볶이에 만두하고, 달걀은 안 넣을거야?|disaster_trial
disaster_trial/wav/disaster_trial_D00001_A00013.wav|많이 먹어서,|disaster_trial	disaster_trial/wav/disaster_trial_D00001_A00013.wav|많이 먹어서,|disaster_trial
disaster_trial/wav/disaster_trial_D00001_A00014.wav|자격지심과, 누구에게랄 것도 없이	disaster_trial/wav/disaster_trial_D00001_A00014.wav|자격지심과, 누구에게랄 것도 없이 대상도 분명치 않는 분노같은

3. 데이터 구축

3.4. 음성 합성 학습용 데이터 구축 과정

• 데이터 구축 과정