

Preprocessing KSS Data (Detail)

≡ 태그

1 Download KSS

[압축 해제](#)

2 Directory

[.wav 파일](#)

[.txt 파일](#)

3 script 편집

① transcript.v.1.4.txt을 Excel로 열기

- 1) 구분 기호로 분리됨 선택 → 다음
- 2) 구분 기호 : 기타 선택 → | 입력 → 다음
- 3) 열 데이터 서식 : 일반 선택 → 다음

② 파일 편집

- 1) 필요 없는 열 삭제 : C, D, E, F 열 삭제

③ 데이터 나누기 : 1 / 2 / 3 / 4

- 1) A열 선택
- 2) 데이터 탭 → 필터
- 3) 필터 드롭다운 클릭
- 4) 텍스트 필터 → 시작 문자
- 5) 시작 문자에 1/1 입력 → 확인
- 6) 본문 내용 전체 선택 → 복사
- 7) 새로운 sheet 생성 후 붙여넣기
- 8) 새로운 sheet 내용 전체 선택 → 복사
- 9) 메모장에 붙여넣기
- 10) 편집 탭 → 바꾸기(R)...
- 11) 찾을 내용에 공백 복사해서 붙여넣기
- 12) 바꿀 내용 : | 입력 → 모두 바꾸기(A) → X
- 13) 메모장 저장
- 14) 5) 부터 13) 반복 : 2/2, 3/3, 4/4
- 15) 최종 생성 파일 : 4개

④ Directory

4 Preprocess Data

[가상환경 활성화](#)

[경로 이동](#)

[preprocess](#)

4개 : 1, 2, 3, 4

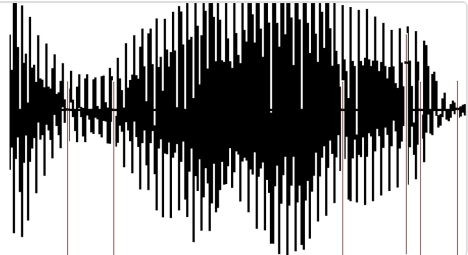
2개 : 1, 2

1 Download KSS

Korean Single Speaker Speech Dataset

KSS Dataset: Korean Single Speaker Speech Dataset

[k https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset](https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset)



- Kaggle에 로그인 후 다운로드
- Archive.zip 다운로드
- Archive.zip 구조
 - kss
 - 1
 - 2
 - 3
 - 4
 - transcript.v.1.4.txt

압축 해제

```
unzip archive.zip
```

2 Directory

.wav 파일

- `fodong/vits/1/`
- `fodong/vits/2/`
- `fodong/vits/3/`
- `fodong/vits/4/`

.txt 파일

- `fodong/vits/filelists/transcript.v.1.4.txt`

3 script 편집

Excel 활용

① transcript.v.1.4.txt를 Excel로 열기

1) 구분 기호로 분리됨 선택 → 다음

텍스트 마법사 - 3단계 중 1단계

데이터가 구분 기호로 분리됨(으)로 설정되어 있습니다.
데이터 형식이 올바르게 선택되었다면 [다음] 단추를 누르고, 아닐 경우 적절하게 선택하십시오.

원본 데이터 형식

원본 데이터의 파일 유형을 선택하십시오.

☒ 구분 기호로 분리됨(D) - 각 필드가 쉼표나 탭과 같은 문자로 나누어져 있습니다.
☐ 너비가 일정함(W) - 각 필드가 일정한 너비로 정렬되어 있습니다.

구분 시작 행(R): 원본 파일(O):

☐ 내 데이터에 머리글 표시(M)

C:\Users\WSSAFY\Desktop\kss_script.txt 파일 미리 보기

1	3/3_0000.wav	아직 만날 사람이 한 명 더 있어.	아직 만날 사람이 한 명 더 있어.	????? ??????
2	3/3_0001.wav	체스에서 컴퓨터가 인간을 이겼어요.	체스에서 컴퓨터가 인간을 이겼어요.	?????????
3	3/3_0002.wav	다섯 명	다섯 명	????? ???2ofive people
4	3/3_0003.wav	세 분	세 분	??? ???1.7three persons
5	3/3_0004.wav	'김'은 한국에서 가장 흔한 성이다.	김은 한국에서 가장 흔한 성이다.	??????? ?????
6	3/3_0005.wav	아기 이름은 수현이라고 지었어요.	아기 이름은 수현이라고 지었어요.	????? ????????

< >

취소 < 뒤로(B) 다음(N) > 마침(F)

2) 구분 기호 : 기타 선택 → | 입력 → 다음

텍스트 마법사 - 3단계 중 3단계

각 열을 선택하여 데이터 서식을 지정합니다.

열 데이터 서식

☒ 일반(G)

☐ 텍스트(T)

☐ 날짜(D): 년월일

☐ 열 가져오지 않음(건너뛰기)(I)

[일반]을 선택하면 숫자 값은 숫자로, 날짜 값은 날짜로, 모든 나머지 값은 텍스트로 변환됩니다.

고급(A)...

데이터 미리 보기(P)

일반	일반	일반
1/1_0000.wav	그는 괜찮은 척하려고 애쓰는 것 같았다.	그는 괜찮은 척하려고 애쓰는 것 같았
1/1_0001.wav	그녀의 사랑을 얻기 위해 애썼지만 헛수고였다.	그녀의 사랑을 얻기 위해 애썼지만 헛
1/1_0002.wav	용돈을 아껴 써라.	용돈을 아껴 써라.
1/1_0003.wav	그는 아내를 많이 아낀다.	그는 아내를 많이 아낀다.
1/1_0004.wav	그 애 전화번호 알아?	그 애 전화번호 알아?
1/1_0005.wav	차에 대해 잘 아세요?	차에 대해 잘 아세요?

< >

취소 < 뒤로(B) 다음(N) > 마침(F)

② 파일 편집

1) 필요 없는 열 삭제 : C, D, E, F 열 삭제

[삭제 전]

transcript.v.1.4.txt - Excel

파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 수행할 작업을 알려 주세요. 로그인 공유

붙여넣기 클립보드 글꼴 맞춤 표시 형식 서식 스타일 삽입 삭제 서식 정렬 및 찾기 및 필터 편집

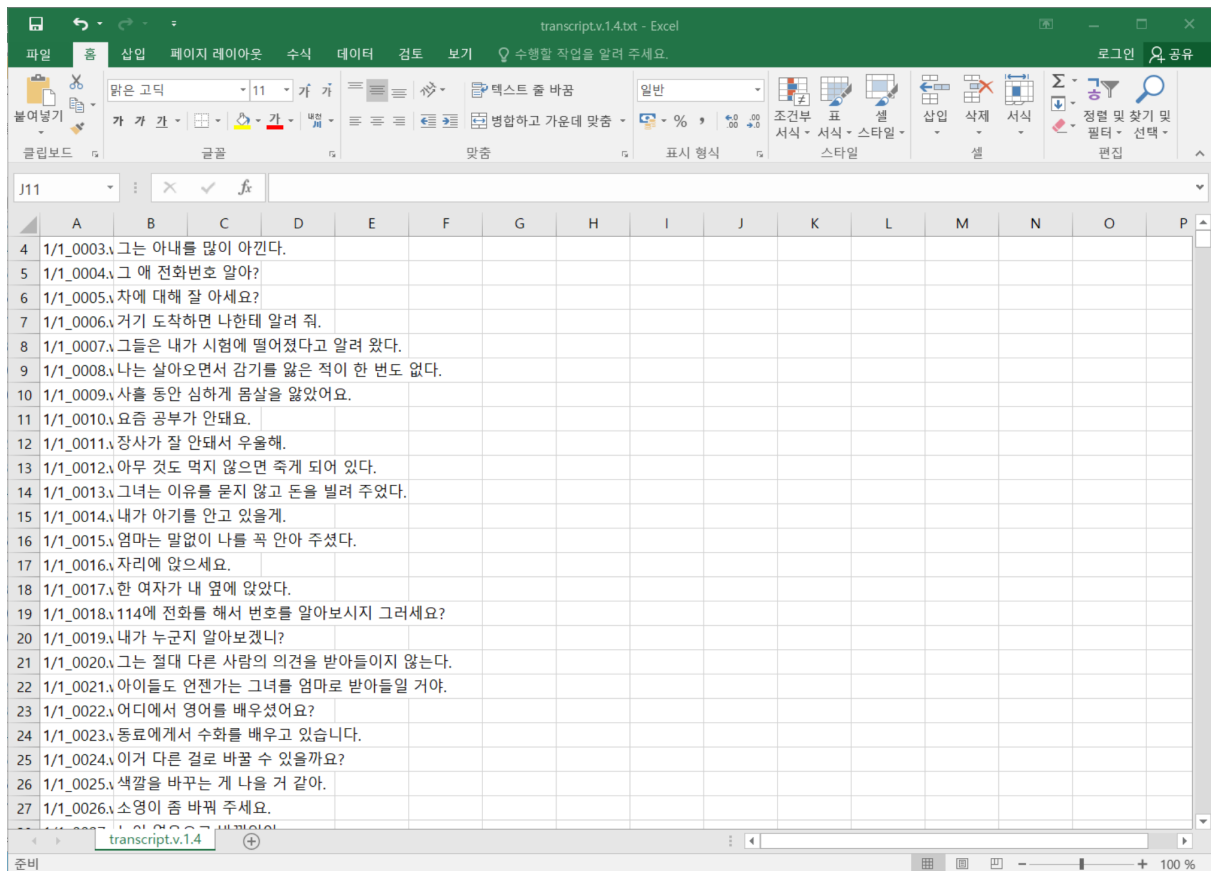
A1 1/1_0000.wav

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
4	1/1_0003.v	그는 아내	그는 아내	2.3	He cherishes his wife.											
5	1/1_0004.v	그 애 전화	그 애 전화	1.3	Do you know his number?											
6	1/1_0005.v	차에 대해	차에 대해	1.7	Do you know much about cars?											
7	1/1_0006.v	거기 도착	거기 도착	2.8	Let me know if you get there.											
8	1/1_0007.v	그들은 내	그들은 내	3.5	They informed me that I failed the exam.											
9	1/1_0008.v	나는 살아	나는 살아	4.2	I've never had a cold in my life.											
10	1/1_0009.v	사흘 동안	사흘 동안	3.2	My whole body ached for three days.											
11	1/1_0010.v	요즘 공부	요즘 공부	1.9	My studying is not going well these days.											
12	1/1_0011.v	장사가 잘	장사가 잘	2.5	I'm depressed because my business is not doing so well.											
13	1/1_0012.v	아무 것도	아무 것도	3.1	If you don't eat anything, you will die.											
14	1/1_0013.v	그녀는 이	그녀는 이	3.8	She lent her money without asking why.											
15	1/1_0014.v	내가 아기	내가 아기	2.2	Let me hold the baby.											
16	1/1_0015.v	엄마는 말	엄마는 말	4	She hugged me tight without a word.											
17	1/1_0016.v	자리에 앉	자리에 앉	1.1	Take a seat, please.											
18	1/1_0017.v	한 여자가	한 여자가	2.2	A lady sat next to me.											
19	1/1_0018.v	114에 전	114에 전	3.8	Why don't you dial 114 and ask for the number?											
20	1/1_0019.v	내가 누군	내가 누군	1.9	Do you recognize me?											
21	1/1_0020.v	그는 절대	그는 절대	4.1	He never accepts the opinions of others.											
22	1/1_0021.v	아이들도	아이들도	4	The children will also accept her as their mom some time.											
23	1/1_0022.v	어디에서	어디에서	2.2	Where did you learn English?											
24	1/1_0023.v	동료에게	동료에게	2.9	I'm learning sign language from my colleague.											
25	1/1_0024.v	이거 다른	이거 다른	2.4	Can I exchange this for another one?											
26	1/1_0025.v	색깔을 바	색깔을 바	2.3	It would be better to change the color.											
27	1/1_0026.v	소영이 좀	소영이 좀	1.8	Can I talk to Soyoung?											

transcript.v.1.4

준비 100 %

[삭제 후]

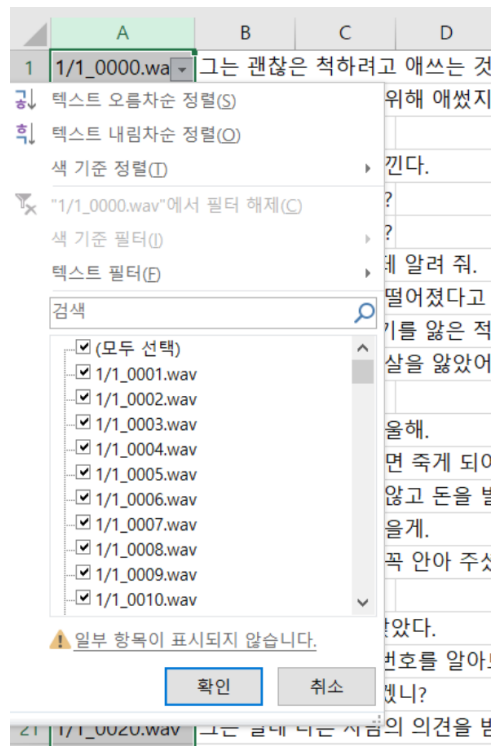


③ 데이터 나누기 : 1 / 2 / 3 / 4

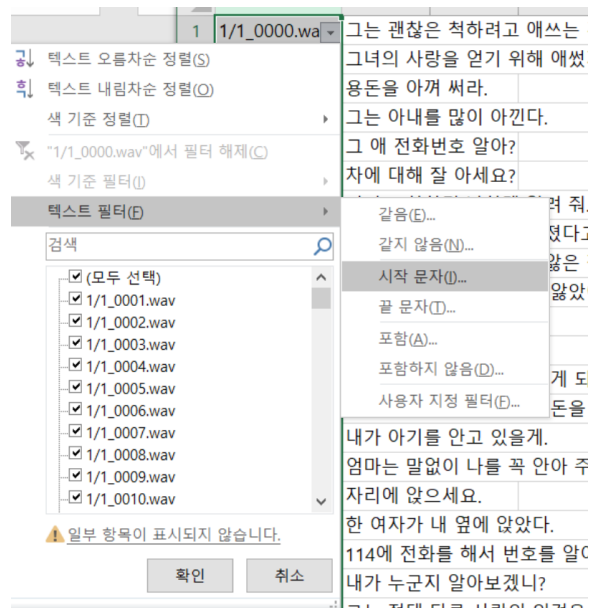
1) A열 선택

2) 데이터 탭 → 필터

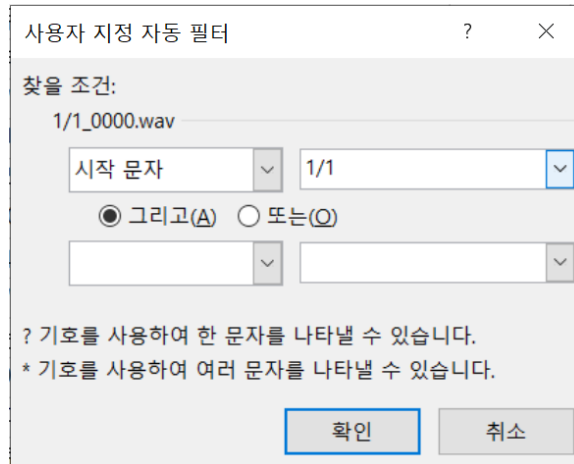
3) 필터 드롭다운 클릭



4) 텍스트 필터 → 시작 문자



5) 시작 문자에 1/1 입력 → 확인

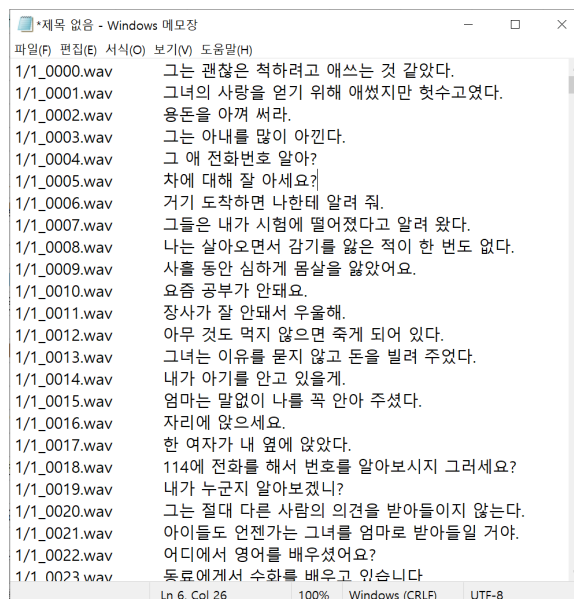


6) 본문 내용 전체 선택 → 복사

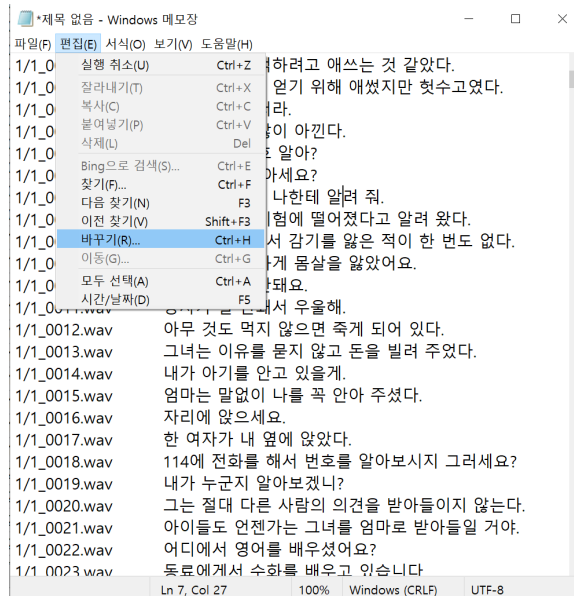
7) 새로운 sheet 생성 후 붙여넣기

8) 새로운 sheet 내용 전체 선택 → 복사

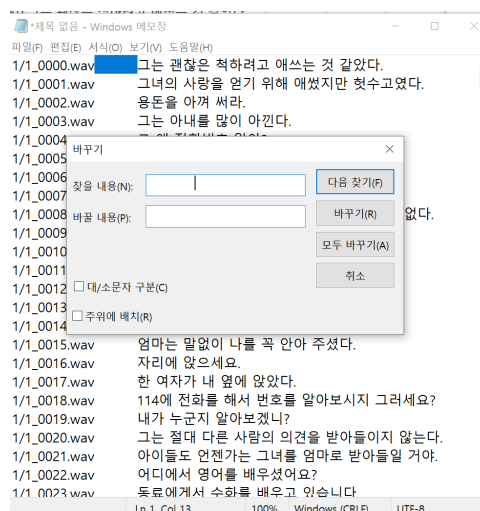
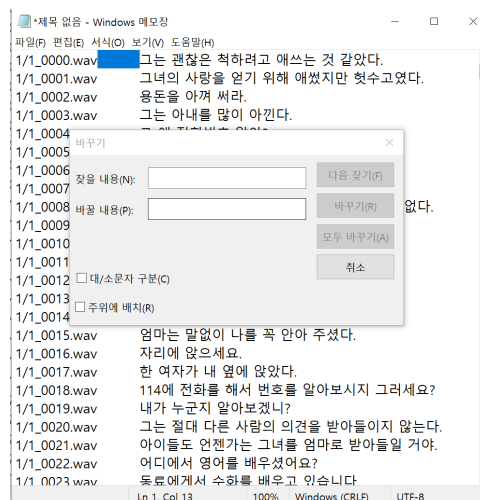
9) 메모장에 붙여넣기



10) 편집 탭 → 바꾸기(R)...



11) 찾을 내용에 공백 복사해서 붙여넣기



12) 바꿀 내용 : | 입력 → 모두 바꾸기(A) → X

13) 메모장 저장

- 파일 명 : kss_script_1.txt

14) 5) 부터 13) 반복 : 2/2, 3/3, 4/4

- 주의할 점
 - 7) 새로운 sheet에 붙여넣기 한 후 첫 번째 행이 1/1이면 해당 행 삭제하기

15) 최종 생성 파일 : 4개

- kss_script_1.txt
- kss_script_2.txt
- kss_script_3.txt
- kss_script_4.txt

④ Directory

- fodong/vits/filelists/kss_script_1.txt
- fodong/vits/filelists/kss_script_2.txt
- fodong/vits/filelists/kss_script_3.txt
- fodong/vits/filelists/kss_script_4.txt

4 Preprocess Data

가상환경 활성화

```
source activate fodong
```

경로 이동

- 경로 : `fodong/vits/`
- `preprocess.py` 파일 확인

```
(base) ljh1004@S220:~/TTSTEST/TTS/vits$ ls
1 LICENSE      configs      kss          modules.py   resources    transforms.py
2 README.md    data_utils.py losses.py     monotonic_align text         utils.py
3 attentions.py filelists    mel_processing.py preprocess.py train.py
4 commons.py   inference.ipynb models.py    requirements.txt train_ms.py
```

preprocess

4개 : 1, 2, 3, 4

```
python preprocess.py --text_index 1 --filelists filelists/kss.
```

```
(tts) ljh1004@S220:~/TTSTEST/TTS/vits$ python preprocess.py --text_index 1 --filelists filelists/kss_script_1.txt filelist
s/kss_script_2.txt filelists/kss_script_3.txt filelists/kss_script_4.txt
START: filelists/kss_script_1.txt
START: filelists/kss_script_2.txt
START: filelists/kss_script_3.txt
START: filelists/kss_script_4.txt
```

2개 : 1, 2

```
python preprocess.py --text_index 1 --filelists filelists/kss.
```

```
(fodong) j-i10c109@jupyter02:~/fodong/vits$ python preprocess.py --text_index 1 --file
lists filelists/kss_script_1.txt filelists/kss_script_2.txt
START: filelists/kss_script_1.txt
START: filelists/kss_script_2.txt
```