한국어 음성 데이터 구축을 위한 전사 작업 가이드라인

한국전자통신연구원, 복합지능연구실
박기영

2024년 1월 18일

차 례

1 개요

이 문서는 한국어 음성 데이터 구축을 위해 전사 작업을 하는 과정에서 필요한 규칙들을 정리한 문서이다. 이 문서는 수시로 변경될 수 있으며 최종 버전은 http://github.com/etri/kmsav/tree/main/trans_guide에서 볼 수 있다.

전사란 말소리를 글로 옮겨 적는 것을 말하며, 특히 본 문서에서는 음성 인식 또는 음성 합성 등 음성 신호와 관련된 기계 학습에 사용할 목적으로, 전사를 하는 방법에 관하여 설명한다.

본 문서에서 의미하는 전사는 아래의 내용을 포함한다.

- 문장 또는 발화 단위로 시간 정보를 기록하는 것.
- 문장 또는 발화 단위로 화자의 ID를 기록하는 것.
- 말소리를 글자로 옮기는 것.
- 말소리 이외에 화자의 간투어, 주변의 잡음, 청자의 호응어 등 특수한 정보를 기록하는 것

과거에는 전사하는 방법으로 “ETRI 전사규칙” 등의 문서가 사용되었다. 이러한 문서에서는 말소리를 글로 옮기는 방법을 **철자전사**, **발음전사**, **표기전사**, **한글전사** 등으로 구분하였다. 각 전사 방법의 의미는 다음과 같다.

철자전사 단어를 표기하는 방식대로 적는 것. 예) KBS1

발음전사 단어를 소리나는 방식대로 적는 것. 예) 케이비에스 완

표기전사 단어를 표기하는 방식대로 적는 것. 철자전사와 같다. 예) KBS1

한글전사 단어를 표기하는 방식대로 적되, 한글로 옮겨서 적는 것. 예) 케이비에스 원

기존의 전사 방식은 각 전사 방식으로 옮겨적는 것이 다를 경우 이 중 두 가지를 병행하는 이중전사를 권장하였다.

(컴퓨터)/(컴퓨타)
(안녕하세요)/(안녕하세여)
(아이폰4)/(아이폰포)
(1001 안경원)/(일공공일 안경원)
(37살)/(서른일곱살)
(1mm)/(일 밀리미터)
(1kg)/(일 킬로그램)
(1/3)/(삼 분의 일)

이러한 이중전사 방식은 과거 음성인식 기술이 음향 모델과 언어 모델로 구분되어 구성되던 시기에, 음향 모델을 학습하기 위해서는 발음전사가, 언어모델을 학습하기 위해서는 한글전사가 유리했던 점을 반영하기 위한 것이었다. 하지만, 음성 인식 및 합성 방법이 종단형 기술로 바뀌면서, 음향 모델과 언어 모델을 구분하여 학습하는 절차가 사라지면서, 이러한 이중전사의 필요성 또한 현저히 줄어들게 되었다. 이에 따라 전사 편의성을 위하여 이중전사를 없애고, 철자전사 및 표기전사 위주로 전사하는 것을 원칙으로 전사규칙을 개정하게 되었다.

2 전사 정보의 저장

전사 정보는 json 형태의 파일로 저장한다. json 파일의 형태는 필요에 따라 추가적인 항목을 가질 수 있으나, 다음 항목은 필수적으로 포함해야한다.

Field	설명
speaker_id	화자의 고유한 ID 정보
start_time	문장의 시작 시간. 소수점을 포함하는 초 단위 숫자로 기록한다.
end_time	문장의 끝 시간. 소수점을 포함하는 초 단위의 숫자로 기록한다.
text	발성 내용 및 기타정보

화자의 고유한 ID 정보는 PKY01과 같이 화자에 부여된 고유한 ID를 사용하거나, A, B, C 등으로 현재 파일 내에서 구분되는 알파벳을 사용한다.

3 화자정보의 전사

여러 화자가 함께 대화한 내용을 녹음하여 전사하는 경우, 분할된 개별 문장에 대하여 화자 정보를 기록한다. 화자 ID가 정해져있으면 해당 ID를 사용하고, 화자 ID가 없는 경우, A, B, C 와 같이 영어 알파벳 대문자 1개로 표시한다. 하나의 대화에서 같은 ID는 같은 화자를 가리켜야한다. 다른 대화에 대해서는 ID가 같더라도, 화자가 다를 수 있다.

4 시간정보의 전사

시간 정보는 문장 단위로 각 문장의 시작과 끝 지점을 파일의 맨 처음을 0으로 하여 초 단위로 기록한다.

- 모든 발화에는 발화의 타임스탬프(시작시간, 끝시간)를 표시한다.
- 시간 단위는 초를 사용하며, 0.1초 이내의 정밀도를 가지도록 전사한다.
- 발화 구간에 대해서만 시간 정보를 기록하고, 그 외의 잡음 및 무음 구간은 기록하지 않는다.
- 발화가 겹치는 구간(오버랩)은 각 발화자의 발화를 개별 발화로 간주하여, 시작 시간 및 끝시간 상으로 겹치도록 전사하며 오버랩에 대한 별도 표기는 하지 않는다. [1]

5 발성내용의 전사

- 발성 내용 전사는 실제 발성한 내용을 그대로 글자로 옮겨 적는 것을 원칙으로 하며, 그 단위는 문장 단위로 한다.

> 사실은 제가 시장님께 궁금한 게 참 많았는데요.
> 네.
> 우리 부산 시민들도 궁금한 점이 참 많더라구요.

[1] 일부 전사문의 경우 /o로 표기한 경우도 있다.

- 각 문장별로 화자 정보를 문두에 표기하는 것도 가능하다.

> [A]사실은 제가 시장님께 궁금한 게 참 많았는데요.
> [B]네.
> [A]우리 부산 시민들도 궁금한 점이 참 많더라구요.

- 이중전사는 하지 않고 표기전사로만 수행한다.

잘못된 전사	바른 전사
(30km)/(삼십 킬로미터)	30 km

5.1 띄어쓰기

띄어쓰기는 표준어법에 맞추어 하되 표준어법으로 명확히 결정할 수 없는 경우에는 띄어 쓴다. 띄어쓰기 공백이 전각 공백문자('\u3000'), 탭 문자 등의 잘못된 문자로 표시되지 않도록 주의한다.

숫자와 이어지는 단어는 붙여쓰는 것을 원칙으로 하되, 이어지는 단어가 수량을 나타내는 단위일 경우에는 한 칸을 띄운다.

발성 내용	바른 전사
삼십 킬로미터	30 km
삼분의이가 지났다	2/3이 지났다
오월에는 다섯개	5월에는 5 개
삼 사십 대	3, 40대
이 삼 천원	2, 3천 원
오뉴월과 칠팔월	5, 6월과 7, 8월
사 오십 프로	4, 50%

5.2 문장부호

문장 부호에 대한 전사규칙은 마침표(.), 쉼표(,), 느낌표(!), 물음표(?) 네 가지 문장부호만을 규정한다. 다른 문장부호는 넣지 않는 것을 원칙으로 한다.

한 문장의 끝나면 이 네개의 문장 부호 중 하나를 반드시 붙여준다. 문장이 종료되지 않고 발화가 종료된 경우에는 쉼표를 넣는다. 쉼표는 한 문장 중간에 필요한

경우에 넣어 줄 수 있다. 문장부호 뒤에는 반드시 공백을 넣어준다.

5.3 그밖의 기호

- 온점은 마침표 이외에 소수점 및 날짜를 나타내거나, 영어 약자를 나타내기 위해 사용되는 경우가 있다. 이 경우에는 온점 뒤에 공백없이 반드시 붙여쓴다. 즉, 온점 뒤에 공백이 있는 경우는 온점이 문장의 마침표인 경우 뿐인다.

발성 내용	전사
삼쩜일사일오구	3.14159
육이오 전쟁	6.25 전쟁
알오케이	R.O.K
씨오쩜 케이알	co.kr

- 슬래시(/) 는 분수를 나타내거나, 영어 약어를 나타내기 위해 사용될 수 있다.

발성 내용	전사
오 분의 삼이 지났다	3/5이 지났다.
에이에스를 받으려고 해	A/S를 받으려고 해.

- +, -, @, & 등 일상적으로 사용되는 기호는 그대로 전사한다.

발성 내용	전사
마이너스 5도	-5도
케이티앤지	KT&G
에비씨앳구글닷컴	abc@google.com

- 긴 발성을 표시하기 위한 표시(~), 말줄임표 표시(…)는 사용하지 않는다.
- α, π 등의 그리스문자나 그밖의 특수기호는 한글로 전사한다.

발성 내용	전사
알파 곱하기 베타 더하기 감마	알파 곱하기 베타 + 감마
이 파이 나누기 삼	2 파이 나누기 3

5.4 방언(사투리)

방언은 표준어로 고쳐적는 것을 원칙으로 하되, 그 특성을 살려야하는 방언의 경우 그대로 표기한다.

예 했어요를 했어유(충청도 사투리)로 발음한 경우, 했어요으로 전사.

예 어머니를 어무이로 발음한 경우, 어머니로 전사.

예 어머니를 어멍(제주도 사투리)로 발음한 경우, 어멍으로 전사.

5.5 영어

- 영어의 경우에는 '외래어' 라면 한국어로, '외국어' 라면 영어로 표기한다. 또한, 일상적으로 영어로 표기하는 것은 영어로 하고, 일상적으로 한글로 표기하는 것은 한글 표기로 한다.

잘못된 전사	바른 전사
youtube	유튜브
wikipedia	위키피디아
white day	화이트데이
wife	와이프
wild food	와일드 푸드
win win	윈윈
won dollar	원 달러

- 또한 영어로 긴 문장을 말하는 경우, 한글식 발음으로 발성했다면, 한국어 음성인식을 위해서는 이는 한글로 전사되는 것이 바람직하다. 만약 영어식 발음으로 발성했다면, 이 발화는 한국어 음성 데이터로는 부적합하므로 전사하지 않고, 데이터에서 제외한다.

잘못된 전사	바른 전사
widespread vaccination	와이드 스프레드 백시네이션
millions of good-paying jobs	밀리언스 오브 굿 페잉 잡스

- 일상적으로 영어로 표기되는 것은 영어로 표기한다.

발성 내용	바른 전사	잘못된 전사
엠비씨	MBC	엠비씨
에스엔에스	SNS	에스엔에스
엘지전자	LG전자	엘지전자
에스케이 하이닉스	SK하이닉스	에스케이 하이닉스 / SK Hynix

- 영어로 읽은 숫자나 기호는 원래의 숫자나 기호로 표기한다.

발성 내용	바른 전사	잘못된 전사
플랜 투	플랜2	plan 2 / 플랜 투
시즌 투	시즌2	season 2 / 시즌 투
투 쓰리	2, 3	two, thre / 투, 쓰리
플러스 포	+4	plus 4 / 플러스 4 / 플러스 포
플러스 마이너스 삼	+-3	plus minus 3

5.6 숫자

- 숫자 표기는 숫자를 살리되, 조, 억과 같은 숫자의 단위는 한글로 살려서 표기한다. 숫자 사이의 콤마는 적는 것이 자연스러운 경우에는 표기한다.

발성 내용	바른 전사	잘못된 전사
백억	100억	10000000000
이천억 개	2,000억 개	200000000000개
일조원	1조 원	1000000000000원
일경 달라	1경 달러	일경 dollar
이백 삼조 오천 사억 원	203조 5,004억 원	203,500,400,000,000원

- 천, 백, 십의 단위도 조, 억 등의 단위와 함께 사용되지 않고, 단독으로 사용되는 경우에는 한글로 살려서 표기한다.

발성 내용	바른 전사
오천원	5 천원
이백개	2 백개
오천이백원	5,200 원

- 숫자와 이어지는 단어의 띄어쓰기는 5.1절을 참고한다.
- **우리말 계열의 수**는 한글로 표기하는 것을 원칙으로 한다. 단 의미하는 수가 10을 초과하는 경우 가독성을 위하여 숫자로 표기한다.

발성 내용	바른 전사
한 달 두 달 세달	한 달, 두 달, 세 달
술 석 잔	술 석 잔
열 한 마리와 열 명	11 마리와 열 명
마흔 아홉명의 사람들	49 명의 사람들

- 시간을 나타내는 경우에는 우리말 계열의 숫자라 하더라도 숫자로 표기한다.

발성 내용	바른 전사
두시 십구분	2시 19분
세시 반	3시 반

- **한자어 계열의 수**는 숫자로 표기하는 것을 원칙으로 한다. 단 **아기돼지 삼형제**와 같이 고유한 명칭이거나, 관용적으로 한글로 쓰는 표현은 한글로 표기한다.

발성 내용	바른 전사
일번문제 한번 풀어봐	1번 문제 한 번 풀어봐
독수리 오형제	독수리 오형제

- 불투명 수량사(서너 개, 사오십, 오뉴월, ··· 등)도 마찬가지로 10 이하의 우리말 숫자에 대해서는 한글로, 그렇지 않은 경우에는 숫자로 표기한다.

발성 내용	바른 전사	잘못된 전사
세 네 개	서너 개	3, 4개 / 3 4 개 / 3-4개
삼 사십 대	3, 40대	3 40대
이 삼 천원	2, 3천 원	2-3천원
오뉴월과 칠팔월	5, 6월과 7, 8월	5 6 월과 7 8 월
사 오십 프로	4, 50%	4 50 퍼센트

5.7 단위

- 영문으로 표기 가능한 것은 영문으로 전사하고, 숫자와 단위는 가급적 한 칸 띄어쓴다.

발성 내용	바른 전사
이리터는 이천미리리터이다	2 L는 2,000 mL이다.
백미터 삼십 오 센티미터	100 m, 35 cm
백팔십센치미터 칠십오킬로그램	180 cm, 75 kg

- 단위가 생략된 경우 발화의 의도가 명백하면, 생략된 기호를 살려서 전사한다. 단 의도를 알 수 없는 경우에는 한글로 전사한다.

오 키로나 빠졌어	5 kg이나 빠졌어
3 센치나 자랐어	3 cm나 자랐어
십이 킬로	12 킬로

- 특수기호가 들어가야 되는 단위는 한글로 전사한다.

발성 내용	바른 전사	잘못된 전사
삼 제곱미터	3 제곱미터	3 m^2
사쩜삼세제곱미터	4.3 세제곱미터	4.3 m^3
백팔십도	180 도	180°
삽심육쩜오도씨	36.5 도씨	35.6℃

5.8 간투어

간투어(filler word)란 발성자가 다음 발성을 준비하기 위해서 소요되는 시간을 벌기 위해서 발성하는 것으로 의미가 없는 것을 의미한다.

- 명확히 간투어로 확인되는 경우 간투어 앞에 f/를 붙여 표기한다. [2]
- 간투어는 표준어 규칙대로 띄어쓰며, f/와 간투어 사이는 띄어쓰지 않는다.

> f/음 내가 그래서 있잖아.

[2] 기존에는 간투어 뒤에 /를 붙여 표시하였다. (예. 아/)

- 간투어로는 아래와 같은 것이 있으며, 이외에 다양한 것이 있을 수 있다.

f/아, f/그, f/어, f/음, f/저기, f/저, f/에, f/음.

- 간투어와 유사한 성격으로 발성한 단어라하더라도, 길이가 길거나 강하게 발음한 경우 일반적인 단어로 취급하여 전사한다.

그러니까 오늘 온다는 말이지?

5.9 오발성 및 알아듣기 힘든 발음

오발성 및 알아듣기 힘든 발음은 사용자의 의도를 추정하여 표준어법이나 의도에 맞게 전사한다.

예 어머니를 어머이로 오발성한 경우, 어머니로 전사.

단, 음성이 명확하지 않은 경우 의도를 추정하기도 어렵다면 u/로 표시해준다.

6 추가 정보의 전사

앞서 설명한 내용은 발화자가 발성한 내용을 글로 표기하는 방법에 관한 것이었으나, 발화자가 발성한 내용이 아닌 것도 일부 필요에 따라, 전사에 포함할 수 있다. 예를 들어, 발성 도중에 발생한 환경 잡음, 웃음소리, 상대방의 호응어 등을 포함한다. 단, 상대방의 의미있는 발성은 추가 정보가 아니라 오버랩으로 처리한다.

6.1 잡음

발화 중 환경 또는 다른 사람의 목소리 등으로 인하여 잡음이 발생한 경우 잡음이 발생한 위치에 n/ 로 표기하여 잡음이 있음을 전사한다.

- 잡음은 주변의 웃음소리, 잡음 등을 포함하며 뚜렷이 구분되는 것만을 표기하며, 발성에 비해 크기가 작은 것은 무시해도 좋다.
- 해당하는 잡음이 날 때마다, 잡음이 들린곳에 단어 단위로 n/를 표기한다.

- 발화 중 전체적인 잡음 환경인 경우 문장의 맨 앞에 n/를 표시한다. 발화자 표시가 있는 경우, 발화자 표시 앞에 한다.

```
n/[A]열어보면은 어! 굉장히 질문이 많아요
n/열어보면은 어! 굉장히 질문이 많아요
열어보면은 n/어! n/굉장히 n/질문이 많아요
```

- 발화자의 웃음소리 중 문장 중간에 있는 분명한 웃음소리는 잡음이 아니며 최대한 웃음소리와 일치하는 한글 발음으로 전사하고, 해당 단어 앞에 l/를 붙여서 웃음소리임을 표시한다.

```
어떻게 l/하하 오셨어요
오늘 할일은 내일의 l/크크크 내게 미룬다.
```

- 발화자의 웃음소리 중 문장 사이에 있는 일반적인 작고 짧은 웃음소리는 전사하지 않으며, 무시한다.

6.2 호응어

호응어란 백채널(back channel)에 관한 것으로, 다른 발화자의 말에 덧붙이는 추임새를 의미한다. 호응어는 두 가지 방법으로 표기할 수 있으며 두 가지 방법을 동시에 적용하는 것도 가능하다. 단, 하나의 데이터셋에 대해서는 전체 데이터에 대하여 일관된 방법으로 적용해야한다.

호응어 화자의 발성내용을 독립 표기

호응어의 위치와 내용을 표기하는 방법으로 호응어 자체를 하나의 문장으로 간주하여, 두 발화가 겹쳐서 발화된 것처럼 표기한다. 호응어의 시작점과 끝점에 대한 시간 정보와 발화자 정보를 표기하고, 호응어의 내용도 발성내용에 맞게 전사한다. 단 호응어임을 표기하기 위하여 문장의 맨앞에 b/를 붙인다.

```
b/그렇죠
b/네
b/네네
b/아
b/음
b/저런
```

주 발화자의 발성내용에 병행 표기

호응어의 위치만을 표기하는 방법으로 주 발화자의 전사 내용 중, 다른 화자가 발성한 호응어의 위치를 b/로 표기한다. 단, 문장의 맨앞에 오는 b/는 이 발화 자체가 호응어임을 의미하므로, 주의하도록 한다.

```
오늘 회사에 갔는데 b/ 쉬는 날이었어 b/
b/어 b/저런
```

7 전사 규칙에 사용된 꼬리표

본 전사 규칙에는 영어 소문자 알파벳에 슬래시(/)를 붙여 꼬리표(tag)로 사용하였다. 사용된 꼬리표의 종류는 다음과 같다.

꼬리표	의미	참고
b/	호응어. 단독으로 사용될 경우 해당 위치에 상대방의 호응어가 있거나, 다른 단어의 앞에 붙어있는 경우 그 단어가 호응어임을 나타냄.	6.2
f/	간투어. 이어지는 단어가 간투어 임을 표시.	5.8
l/	웃음소리. 이어지는 단어가 웃음소리임.	6.1
n/	잡음. 단어 앞에 붙은 경우에는 해당 위치에서 환경 잡음이 발생했음을 의미하고, 문장 가장 앞에 붙은 경우에는 발성구간 전체에걸쳐 잡음이 존재함을 의미한다.	6.1
u/	오발성. 알아듣기 힘들고 의도도 알 수 없는 발성.	5.9

8 맺음말

지금까지 종단형 음성인식기의 학습에 적합한 한국어 음성데이터의 전사 규칙에 관하여 설명하였다. 종단형 음성인식기는 그 출력이 사용자에게 그대로 제시되므로, 학습과정에서도 사용자에게 가독성이 높은 형태의 전사문으로 훈련되는 것이 바람직하며, 전사규칙의 원칙도 이에 따른다. 다만 이러한 전사 규칙이 절대적인 것은 아니며, 사용 목적에 따라 적절히 수정하여 사용하는 것이 바람직하다.

연락처

박기영(pkyoung@etri.re.kr)