ETRI 전사규칙

개정이력

- 2019.5.8. 버전 1.0 수립 (음성지능연구그룹)

1.1. 개요

1.1.1. 표준발성에서 벗어나거나 같은 전사에 대하여 두 가지 이상 발음이 가능한 경우 발음전사와 철자전사를 병행하며, 이 경우 (철자전사)/(발음전사)로 표기한다 (이 문서에서 향후 이를 '이중전사'라 칭한다).

예) (컴퓨터)/(컴퓨타)

1.1.2. 발음전사: 발성된 내용을 소리 값에 최대한 가깝게 표기한다. 이는 음성인식의 음향모델링을 주된 목적으로 한다.

1.1.3. 철자전사: 표준어법에 맞게 표기한다. 이는 음성인식의 언어모델링 등을 주된 목적으로 한다.

1.1.4. 숫자, 외래어, 기호, 도량형 및 온도 단위는 발음 전사를 수행하되, 별도의 목록표를 생성하여 발음 전사별로 해당되는 표준 표기를 명시한다 (1.3, 1.7, 1.8절 참조).

1.1.5. 이중전사를 할 때, 이중전사의 범위를 표시하기 위해 괄호('(', ')')를 사용한다.

1.1.6. 이중전사, 잡음, 중복 발성 등을 나타내기 위한 특수 기호(meta symbol, 예: '/', '(', ')', '*', '+')는 원래의 목적으로만 표기되어야 한다. 특수기호가 실제 발성된 경우에는 발성된 형태를 반영하여 발음전사 한다. 분수 표기도 풀어서 표기한다.

1.1.7. 전사 과정에서 삽입되는 모든 기호('(), /'등)는 아스키코드만 사용하도록 한다.

예)

- 1/3 -> 삼 분의 일
- 슬래시, 작대기, slash
- 별표, star sign, asterisk
- 덧셈기호, 더하기, plus

1.1.8. 이중전사 할 때 '/'와 앞 뒤 괄호 사이에는 space를 두지 않는다.

1.2. 잡음

1.2.1. 단어의 앞과 뒤에 거의 붙어 발생된 잡음은 단어와 분리하여 표기한다.

1.2.2. 잡음이 있는 상황에서 사람에게서 발생하는 잡음은 명확히 구분될 정도로 큰 것만 표기해도 좋다.

1.2.3. 다음에 정의된 잡음 이름 뒤에 '/'를 붙여 표기한다.

- b : 숨소리

- l : 웃음 소리(laugh)
- o : 다른 사람의 말소리가 포함된 경우 문장의 맨 앞에 표기
- n : 주변의 잡음

1.3. 숫자 표현

1.3.1. 기본적으로 숫자는 모두 숫자 기호가 아닌 문자로 표현하며, 필요한 경우 별도의 목록표를 작성한다.

- 숫자는 발음한 형태를 반영하여 문자로 표현하며, 한국어 및 영어에 대해 동일하게 적용한다.
- 한국어의 경우 십진 단위로 띄어 쓴다. 숫자를 하나씩 발음한 경우에도 띄어 쓴다.
- 단위를 나타내는 '년', '월', '일', '시', '분' 등은 숫자와 띄어 쓴다.
- 경계 내부에 설령 간투어 또는 잡음 등이 포함되어 있다고 하더라도 포함된 간투어를 포함한 상태로 경계를 표시해 준다.

예)
- (5대)/(오 대) 그룹이 모여, 자동차 (5대)/(다섯 대)를
- (24시간)/(이십 사 시간), (24시간)/(스물 네 시간)
-(867-860-2437)/(팔 육 칠 팔 육 공 에 이 사 삼 칠)
-(14시)/(십 사 시), (14시)/(열 네 시)부터
-(1999년)/(천 구백 구십 구 년)에, (1999년)/(일천 구백 구십 구 년)에

1.3.2. 숫자만으로 이루어진 기념일 등 특정 의미가 있는 단어들을 목록을 별도 작성한다. 이 때, 아라비아 숫자에 붙는 단위, 조사나 어미는 붙인다.

예)

팔 일 오	8.15
사 일 구	4.19
오 칠 오 공 부대	5750부대

1.4. 간투어 표현

1.4.1. 발성자가 다음 발성을 준비하기 위해서 소요되는 시간을 벌기 위해서 발성하는 것으로 의미 없는 것을 말한다. 간투어 뒤에 '/'를 붙여 표기한다.

1.4.2. 예) 아/, 그/, 어/, 그/, 아/, 음/, 저/, 저기/, 에/, 으/, 응/, … 등

1.5. 외국어/외래어/약자

1.5.1. 일반적으로 외국어 문자로 표기하는 경우, 통상의 발음대로 읽은 경우는 통상의 표기를 따른다.

예) KBS, MBC, AT&T, ETRI, OPEC, FIFA 등

1.5.2. 우리말로 표기하여 자연스러운 것은 한글로 표기한다. 애매한 경우도 한글로 표기한다.

예) 뉴욕, 시카고, 파티, 버스, 핸드폰, 모바일, 인터넷, 호텔 등등

1.5.3. 통상적인 발음으로 읽는 외국어/외래어/약자들에 대한 목록표를 별도 작성한다.

1.5.4. 통상적인 발음으로 읽지 않은 경우, 이중 전사한다 (2.8.4절 참조).

1.6. 문장 부호

1.6.1. 문맥적인 의미를 파악하여 표기하며, 한 문장이 끝나면 반드시 문장부호(마침표, 물음표, 느낌표)를 표기하며, 중간에 문맥의 표시를 위해 쉼표 ',’는 허용을 한다.

1.7. 도량형 및 온도, 단위

1.7.1. 온도 등의 단위는 발음을 반영하여 한글/영어 문자로 적어준다.

1.7.2. “Degree Celsius”와 같이 띄어 쓰는 것이 분명한 경우를 제외하고는 붙여 쓴다.
예) kilometer (O), kilo-meter (X), kilo meter (X)

1.7.3. 모든 도량형은 목록표를 별도 작성한다. 이 때, 목록표에는 유로, 프랑 등 키보드에 없는 기호도 포함한다.
예)

밀리미터	mm
미리미터	mm
미리	mm
밀리메터	mm
킬로그램	kg

1.7.4. 숫자, 기호, 영문표기에 대하여 표준발음과 달리 발성된 경우, 한국어는(철자표기)/(실제발음)로 표기한다. 이 경우는 한국어는 전사문 작성자가 귀로 들었을 때 발음 자체는 명확히 들리는 경우이며, 발음 자체가 불명확한 경우는 2.10절을 참고한다.
예) (UNESCO)/(유네코) : '유네스코'를 '유네코'로 잘못 발성한 경우
(UNESCO)/(유 엔 이 에쓰 씨 오) : ‘유네스코’를 ‘유 엔 이 에쓰 씨 오’로 알파벳으로 읽은 경우 (한국어)
(UNESCO)/(U N E S C O) : ‘유네스코’를 ‘유 엔 이 에쓰 씨 오’로 알파벳으로 읽은 경우 (영어)
UNESCO // ‘유네스코’라고 통상의 방식대로 발성한 경우
(다섯)/(다싯) 대 // '다섯 대'를 '다싯 대'로 잘못 발성한 경우

1.8. 띄어쓰기

1.8.1. 띄어쓰기는 표준어법에 맞추어 하되 표준어법으로 명확히 결정할 수 없는 경우에는 띄운다.

1.9. 알아듣기 힘든 발음

1.9.1. 화자가 발음한 내용을 잘 알아 듣기 힘들 때 어절의 뒷부분에 '*'를 붙여 이중전사한다. 즉 전후 문맥을 보고는 알 수 있으나 한 단어만을 놓고 볼 때 발음을 잘못하여 분명히 알 수 없을 때 붙여준다. 명확히 발성된 경우는 '*'를 붙이지 않는다.

예) 나는(이렇게)/(이럴끄*) 그것을 해결하였다. (청취시 '이럴끄'와 비슷하게는 들리지만, 분명히 알 수 없을 때)

나는(이렇게)/(이럴끄) 그것을 해결하였다. (청취시 '이럴끄'가 분명히 들리는 경우)

1.9.2. 방언에 해당하는 발성은 다음과 같이 이중 전사를 한다.

예) (장의사)/(장으사), (학교)/(핵교)

1.9.3. 문맥을 고려해봐도 전혀 알아들을 수 없는 발화는 'u/' 으로 표기한다.

1.9.4. 발성과 동시에 발생하는 잡음은 어절 끝에 '*'를 붙여 표기한다.

예) 기차 타는 곳이* 어디입니까? // '곳이' 가 발성될 때 외부잡음이 크게 섞임

1.9.5. 반복 발성이나 잘못된 발성은 반드시 표기 한다. 이때 불필요하게 중복 또는 잘못 발성된 부분은 뒤에 '+'를 붙인다. 예) 아침에 학교+ 학교에 갔다.

I don't have sta+ stati+ statistical knowledge.

1.9.6. 반복 발성의 발음이 불분명할 때 * 와 + 를 병기한다. 예: "학교*+ 학교에 갔다."

1.9.7. 대화체 문장은 문장 자체가 이상하더라도 그대로 전사한다.