

●○ 자연어 영역

한국어 음성



●○ 개요 : 인공지능학습용 자유발화 음성 데이터셋이란?

딥러닝 기반 음성인식 기술의 급속한 발전과 함께 기존의 낭독체 음성데이터 이외에 AI비서, 로봇, 통역, AI콜센터 등 인간-기계간 자유롭게 대화하는 것을 자연스럽게 인식할 수 있는 기술 개발을 목표로 자유발화 음성데이터의 수요가 점진적으로 높아지고 있다. 이에 본 데이터는 자유발화 음성을 대량으로 수집하여 수집된 음성 발화에 대해서는 전사를 통하여 인공지능학습용 음성데이터를 말하며 국내외 공개되어 연구개발에 적극 활용되길 기대하고 있다.





낭독형 제한발화 음성인식 (As is)	대화형 자유발화 음성인식 (To be)
<ul style="list-style-type: none"> (낭독형) 발음을 명료하게 또박또박 발성 (제한발화) 대본있음  <p><대본을 낭독하여 녹음></p> <ul style="list-style-type: none"> (난이도) 표준규칙에 따라 발성 (응용사례) 단어나 문장단위 제한발화로 스마트폰 음성검색, 덕데이션, 네비게이션 주소검색, 음성리모콘 등에 제한 범위내 활용  <p><음성검색 및 음성명령></p>	<ul style="list-style-type: none"> (대화형) 발음이 불명료하고 유창하게 발성 (자유발화) 대본없음  <p><대본없이 대화상황을 녹음></p> <ul style="list-style-type: none"> (난이도) 간투사, 생략, 더듬거림, 잘못발성, 사투리 포함 (응용사례) 사람들 또는 기계와 자유대화로 동시통역, 인공지능 비서, 대화로봇, AI 콜센터, 대화형 전문가시스템 등 활용 급증 예상  <p><대화형 AI 및 동시통역></p>

그림1 | 기존 낭독체음성과 자유발화 음성과 차이점 비교

●○ 데이터셋의 구성

대상 언어는 한국어이고 구축 분량은 대화음성 1,000시간에 해당한다. 두 사람이 조용한 환경에서 자유롭게 대화하는 음성을 녹음하고 전사규칙(부록: ETRI 전사규칙)에 따라 음성내용이나 잡음을 소리값에 충실하게 일관성 있게 전사한다. 최종 음성데이터 포맷은 16KHz, 16bits headerless linear PCM, 텍스트는 EUC-KR 코드로 저장한다. 본 DB 구축에 참여한 전체 화자는 총 2,000명으로 성별 비율은 다음과 같다. 연령별, 지역별 비율은 특정하지 않았지만 구축기간 등을 고려할 때 대략 상식적인 선에서 분포가 할당되도록 한다.

표1 | 화자 구성

구분	남	여
인원	923명 (46%)	1,077명 (54%)
계	2,000명	

●○ 데이터셋의 설계 기준과 분포

발화자 구성은 성별 비율이 대체로 50% 정도 되도록 고려하고, 현실적으로 어려운 사항이 있는 경우 -5%~+5% 분포 차이는 허용하도록 한다. 한 화자 당 녹음 시간 30분 제한하고, 동일화자 중복 제한을 둔다. 두 사람의 대화 주제는 참여자가 서로 상의하여 자유롭게 선정하게 하거나 관리자가 제시하여 진행하며, 대화 주제가 편중되지 않도록 주의한다.

표2 | 대화 주제 예시

안부 일상 대화	자기 소개	날씨	계절
	거주지 정보		황사/미세먼지
	이성친구		혹서기/혹한기
	학교생활		장마/폭설
	회사생활		온도
	기념일		눈/비/안개 등

쇼핑	의류	취미	사진
	전자기기		여행
	생활용품		음식(맛집)
	악기 등		책
TV	예능		운동
	드라마		전시회
	영화		공연
	연예인		블로그
	시사		음악
정치 경제	정치		스포츠
	부동산		게임
	주식		자동차

음성데이터 전사를 위한 신호처리, 파일링 등 전처리 기준은 다음과 같이 설계한다.

- 음성은 문장 단위로 저장한다.
- 음성파일을 만들 때 반드시 음성구간이 잘리지 않아야 한다.
- 음성구간 앞, 뒤에 200msec 이상 휴지(pause)길이가 포함되어야 한다.
- 잡음이 음성구간의 200msec 내에 포함되어 있는 경우, 잡음을 포함하여 파일로 만들며 잡음 외에 200msec 정도의 휴지 길이가 음성 앞, 뒤에 포함되어야 한다.
- PC마이크, PC헤드셋 환경에서는 16kHz Sampling, 16bit linear PCM으로 저장한다.
- Binary byte order는 Little Endian(Intel 규격)으로 저장한다.
- 클리핑이 발생하지 않도록 음성 크기(amplitude)의 최대값이 16bit 기준 10,000~20,000 사이가 되도록 레코딩 볼륨을 적절히 조절한다.

자유발화 토픽(주제)은 다음과 같이 제한을 두도록 설계한다.

- 두 화자가 높은 자유도로 대화하는 상황을 녹취한다.
- 화자는 스크립트를 가지고 읽을 수 없으며 녹음 상황을 인식하지 않도록 주의를 주어야 한다.
- 다양한 주제로 이루어진 토픽을 이용하여 두 화자가 높은 자유도를 유지한 상태에서 자유발화하며 그 대화를 녹음 수집한다.
- 토픽은 대화에 참여하는 발화자가 선택하도록 하나 중복된 토픽을 피하도록 관리자가 토픽을 제시하거나 유도하도록 한다.

음성데이터 전사 기준은 (부록)과 같이 정한다.

●○ 데이터 구조

본 음성 DB는 훈련용 데이터와 평가용 데이터로 나누어져 있다.

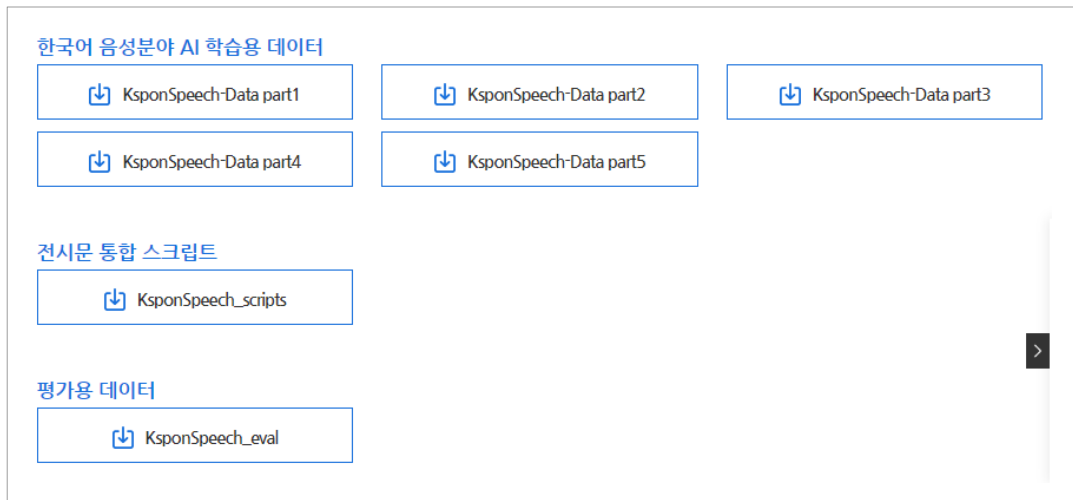
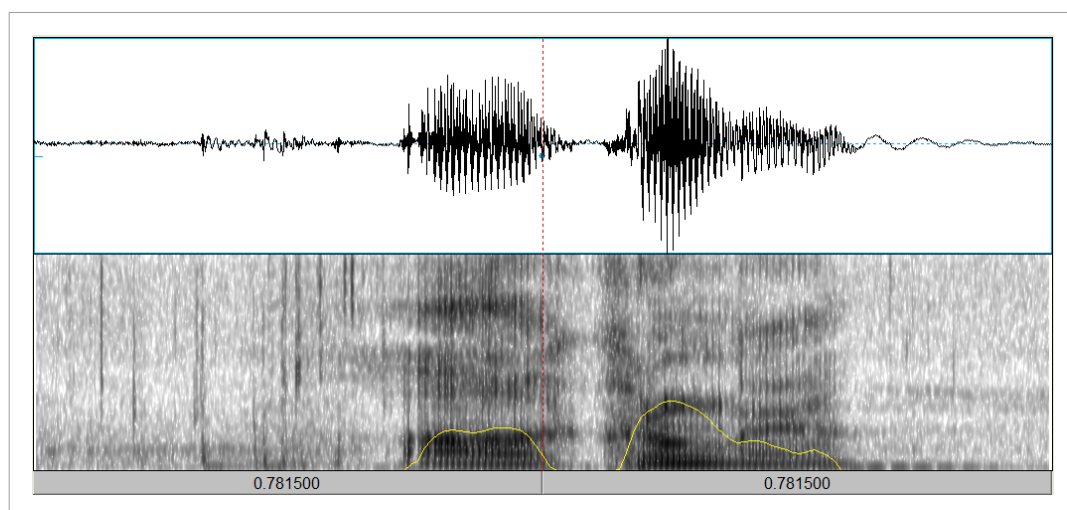


그림2 | AI hub site(<https://aihub.or.kr/aidata/105/download>) 참조

- 학습용 데이터는 /KsponSpeech 폴더 안에 1개 문장에 해당하는 pcm 음성파일과 txt 전사문으로 구성된다.
예) /KsponSpeech_0001/KsponSpeech_000001.pcm, KsponSpeech_000001.txt
- 전사문 통합스크립트는 개발자가 언어모델이나 텍스트 처리 편의를 위해 각 폴더안에 있는 전사문을 통합한 파일이다.
- 평가용 데이터는 학습용 데이터에 사용되지 않은 데이터로 난이도에 따라 Eval_clean (2.6시간), Eval_other(3.8시간)로 나누고, 화자적응용 Train_adapt(25.7시간) 등 크게 3가지로 구성된다.

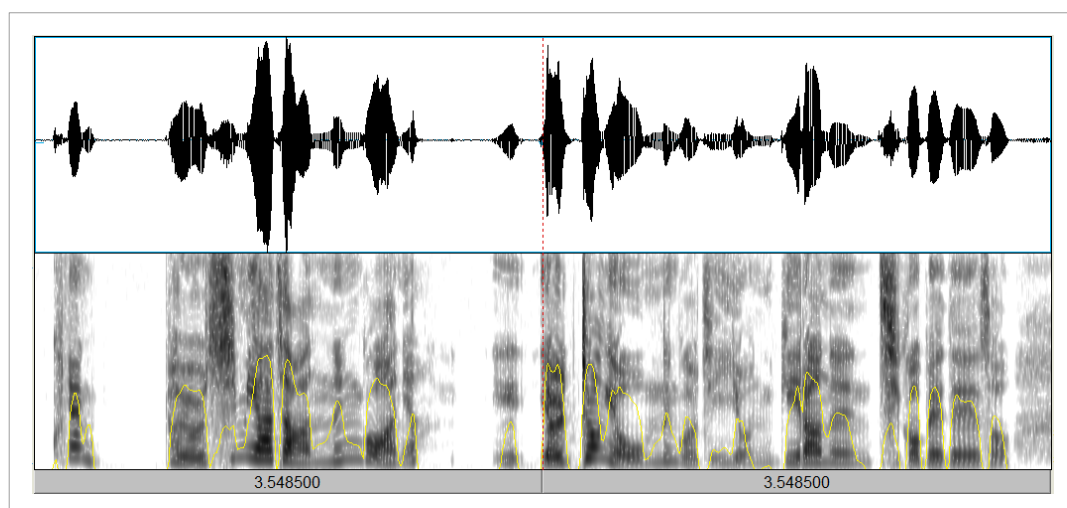
●○ 데이터 예시

다음 그림은 AI hub site에 공개중인 데이터 중 일부 예시이다.



n/ 아! 그런가요

그림3 | KsponSpeech_001001.pcm의 음성파형, 스펙트럼 및 전사문



그냥 별 열심히 하지 않은거 아냐? 이 열정이 없는 거지. 연기라는 직업에 대해서 b/

그림4 | KsponSpeech_001002.pcm의 음성파형, 스펙트럼 및 전사문

●○ 데이터 구축 과정

자유발화 음성데이터 구축 과정은 다음과 같다.

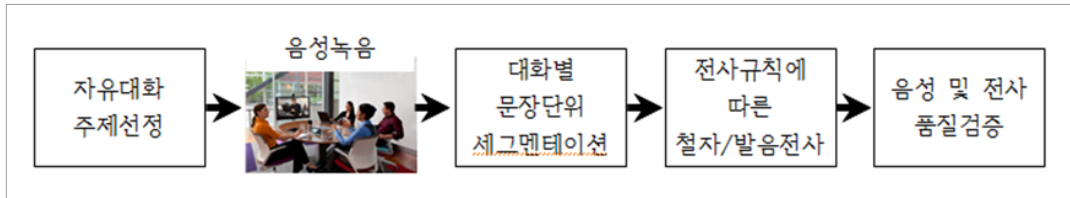


그림5 | 구축 과정

데이터 수집은 조용한 사무실 환경에서 수행하고, 내부에 가구나 흡음재를 배치하여 사무실 공간 특성으로 인한 반향이 발생하지 않도록 한다.

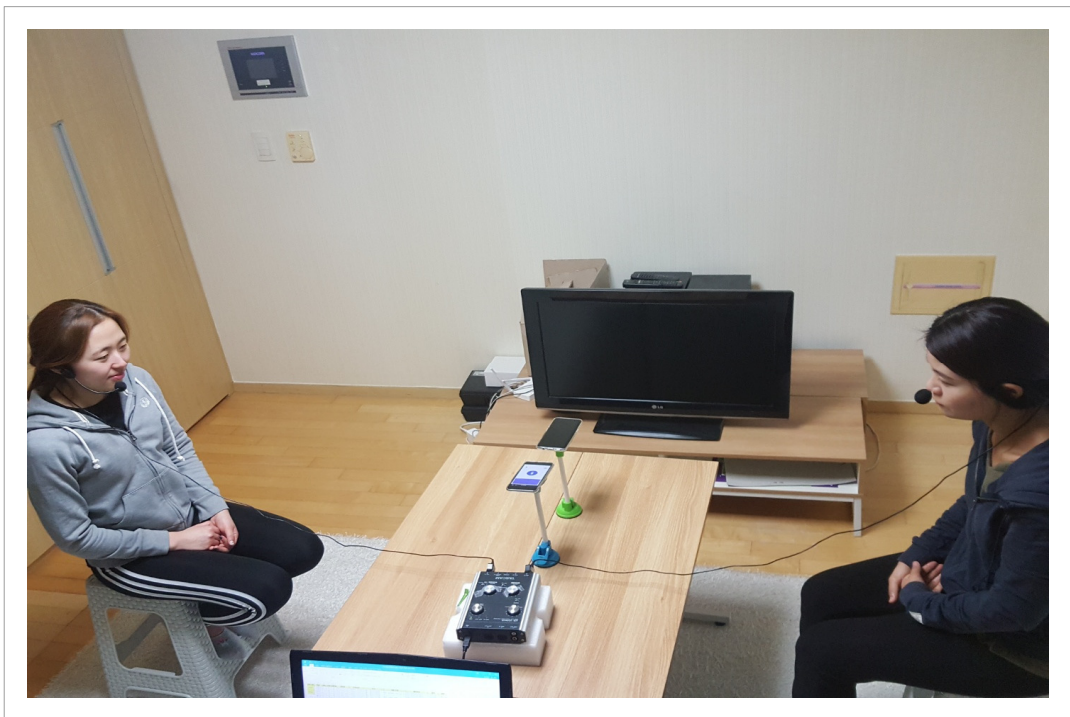


그림6 | 대화 음성 녹음 장면

대화 음성 녹음은 두 사람이 헤드셋을 착용한 상태에서 서로 자연스러운 대화를 하도록 유도하여 녹음이 되도록 구성한다. 일반적으로 두 사람의 자유발화 특성 상 발생이 중첩(Overlap)되는 것을 피할 수 없으나, 헤드셋 마이크의 녹취는 발화간 중첩이 되지 않도록 화자 간 거리를 조절하여 진행한다. 헤드셋 데이터는 오디오 인터페이스를 통해 노트북에 저장되고 사용된 장비는 다음과 같다.

표3 | 마이크와 오디오 녹음 장비 스펙

구분	모델명	특징
마이크	Shure WH20	단일지향성, 다이내믹 마이크
오디오 인터페이스	Tascam US-122	USB2.0 audio interface
녹음 SW	CoolEditPro	안정화된 오디오 편집 도구

녹임데이터는 후처리 과정을 통해 두 화자의 음성을 스테레오로 녹음한 후, 모노로 분리하여 편집과 전사를 진행한다.



그림7 | 2채널 데이터 분리 후

분리된 데이터는 음성데이터 전사를 위한 신호처리 기준, 음성데이터 전사기준에 따라 편집한다.

●○ 검수와 품질 확보

품질확보를 위해 다음과 같은 구축기준, 프로세스, 사용도구 등을 고려한다.

- 일관성 있는 품질확보를 위한 DB구축 기준 마련
 - 대화음성 전사는 주관기관에서 정한 'DB 구축을 위한 전사 가이드라인' 따르되, 예외사항이 있는 경우 주관기관과 협의하여 기준을 보완 적용
 - 사업 수행초기에 주관기관의 요구사항을 충분히 이해한 후 DB구축 작업에 착수하도록 하고, 주관기관은 Critical한 오류가 없는지 재차 확인하고 추진 관리
 - DB 구축 프로세스에 따라 샘플 데이터 구축, 전사 가이드라인에 따라 발음/철자 전사 등 검증 후 본격적인 DB구축 추진
- 효율적 데이터 생성을 위한 DB구축 프로세스 수립
 - 음성녹음, 전사 품질을 제고하기 위한 DB 구축 프로세스가 있어야 하며, 이를 주관기관에 제시하여 승인을 득한 후 전체 프로세스에 적용해야 함

- 사업자는 자연스러운 대화를 유도할 수 있도록 대화주제 선정, 발성환경 셋업 등 주관기관과 협의하여 추진
- 원시 데이터로부터 고품질 DB 확보를 위한 DB화 공정과정 구체화
 - 구축시스템 설계, 원시자료 수집, 데이터 입력, 검증, DB패키지, 홍보, 배포 등 DB생성부터 배포까지 DB화 공정을 수립
 - 기 구축양 보다 110% 초과 구축하여 최종 DB배포에 일정지연이 없도록 추진
- 작업자의 입력오류를 최소화 하기 위한 데이터 전사용 전문도구 사용
 - 전사 도구는 공개 도구인 Transcriber를 사용하였고, 다화자가 발성한 경우에 화자를 구분하여 전사할 수 있는 장점이 있다.

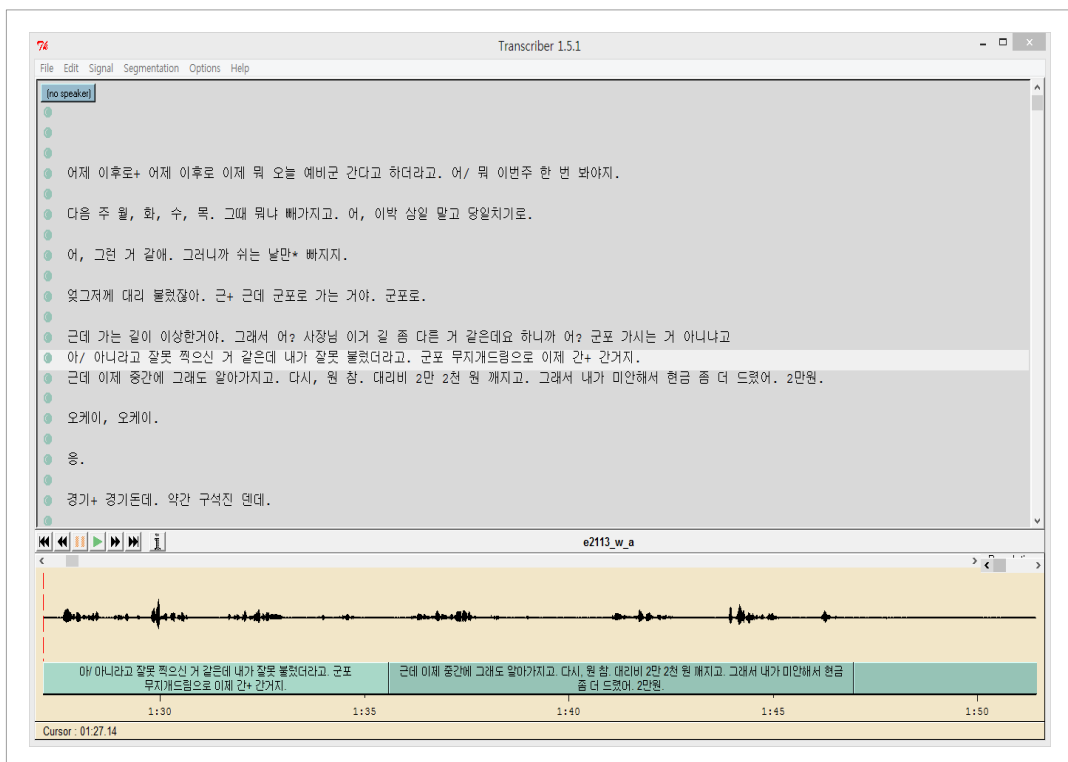


그림8 | 전사 도구(Transcriber)(<http://trans.sourceforge.net/en/presentation.php>)

- 데이터 검사와 관련, 편집 및 전사가 완료된 데이터를 대상으로 음성과 전사문이 일치하는지, 전사문은 전사규칙에 따라 작성되었는지 전수 검사한다.

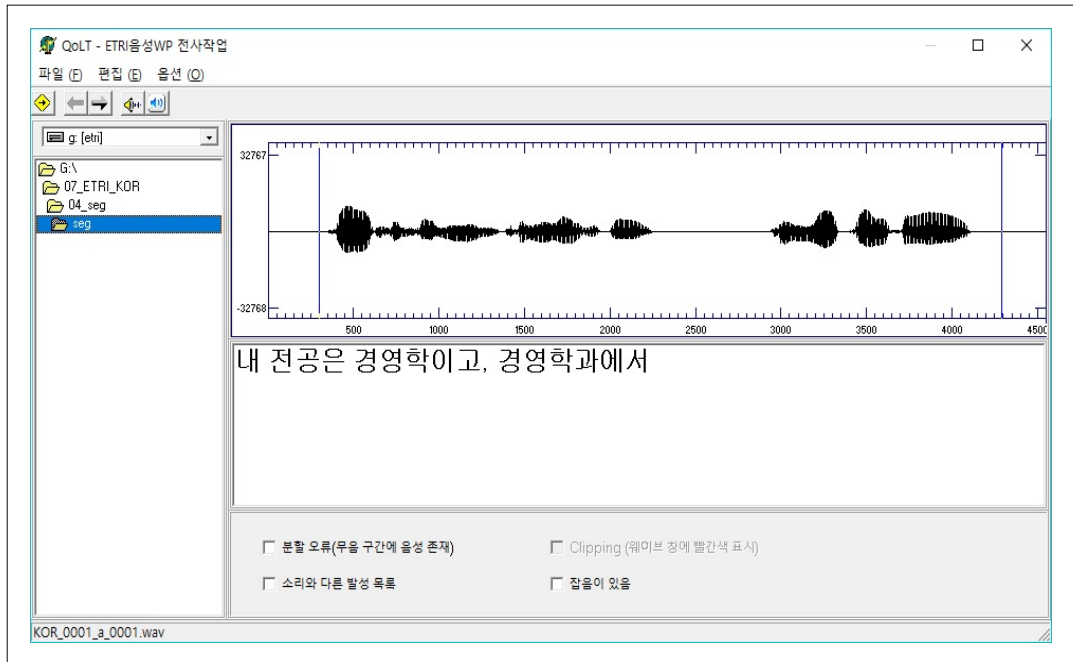


그림9 | 음성 검사 프로그램

●○ 데이터 구축 담당자

수행기관(주관) : ETRI 복합지능연구실 김상훈 책임연구원
(이메일: hello@mindsfab.ai)

●○ 부록

ETRI 전사규칙



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

예) (컴퓨터)/(컴퓨터)

- 2) 발음전사: 발성된 내용을 소리 값에 최대한 가깝게 표기한다. 이는 음성인식의 음향모델링을 주된 목적으로 한다.
- 3) 철자전사: 표준어법에 맞게 표기한다. 이는 음성인식의 언어모델링 등을 주된 목적으로 한다.
- 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).
- 5) 이중전사를 하거나 연속 숫자 등을 전사할 때, 이중전사 또는 연속 숫자 등의 범위를 표시하기 위해 괄호('(', ')', '[', ']')를 사용한다.
- 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.
- 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

- 8) 이중전사 할 때 '/'와 앞 뒤 괄호 사이에는 space를 두지 않는다.

가. 잡음

- 1) 단어의 앞과 뒤에 거의 붙어 발생한 잡음은 단어와 분리하여 표기한다.
- 2) 잡음이 있는 상황에서 사람에게서 발생하는 잡음은 명확히 구분될 정도로 큰 것만 표기해도 좋다.
- 3) 다음에 정의된 잡음 이름 뒤에 '/'를 붙여 표기한다.

lg : 웃음 소리(laugh)

br : 숨소리

n : 주변의 모든 잡음 (음성이 들지 않을 정도의 심한 잡음)

나. 숫자 표현

- 1) 기본적으로 숫자는 모두 숫자기호가 아닌 문자로 표현하며, 필요한 경우 별도의 목록표를 작성한다.
- 2) 숫자는 발음한 형태를 반영하여 문자로 표현하며, 한국어 및 영어에 대해 동일하게 적용한다.
- 3) 연속 숫자로 사용된 경우, 단위와 함께 사용된 경우가 있으므로 일괄 후처리의 편의를 위해 그 경계를 표시해 준다. 경계는 '['와 ']'로 표시한다.
- 4) 한국어의 경우 십진 단위로 띄어 쓴다. 숫자를 하나씩 발음한 경우에도 띄어 쓴다.
- 5) 단위를 나타내는 '년', '월', '일', '시', '분' 등은 숫자와 띄어 쓴다.
- 6) 연속 숫자, 단위를 포함한 숫자와 전화 번호와 같은 경우 그 경계의 시작을 '[', 끝을 ']'을 이용하여 표시해 준다.
- 7) 경계 내부에 설령 간투어 또는 잡음 등이 포함되어 있다고 하더라도 포함된 간투어를 포함한 상태로 경계를 표시해 준다.

예)

- [오 대] 그룹이 모여, 자동차 [다섯 대]를
- [이십 사 시간], [스물 네 시간]
- [팔 육 칠], [팔 육 공 에 이 사 삼 칠]
- [십 사 시], [열 네 시]부터
- [천 구백 구십 구 년]에, [일천 구백 구십 구 년]에
- [Twenty five kilogram], [Three hundred fifteen kilogram]

- 8) 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

다. 간투어 표현

- 1) 발생자가 다음 발성을 준비하기 위해서 소요되는 시간을 벌기 위해서 발생하는 것으로 의미 없는 것을 말한다. 간투어 뒤에 ‘/’를 붙여 표기한다.

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

라. 외국어/외래어/약자

- 1) 일반적으로 외국어 문자로 표기하는 경우, 통상의 발음대로 읽은 경우는 통상의 표기를 따른다.

예) KBS, MBC, AT&T, ETRI, OPEC, FIFA 등

- 2) 우리말로 표기하여 자연스러운 것은 한글로 표기한다. 애매한 경우도 한글로 표기한다.

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

- 3) 통상적인 발음으로 읽는 외국어/외래어/약자들에 대한 목록표를 별도 작성한다.

- 4) 통상적인 발음으로 읽지 않은 경우, 이중 전사한다 (1.8.4절 참조).

마. 문장 부호

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

바. 기호

- 1) 모든 기호는 발음한 형태로 표기하고, 목록표를 별도 작성한다.

예) [이 삼 오 다시 삼 사 칠]

다시	-
슬래시	/

사. 도량형 및 온도, 단위

- 1) 온도 등의 단위는 발음을 반영하여 한글/영어 문자로 적어준다.

- 2) “Degree Celsius”와 같이 띄어 쓰는 것이 분명한 경우를 제외하고는 붙여 쓴다.

예) kilometer (O), kilo-meter (X), kilo meter (X)

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

예)

밀리미터	mm
마리미터	mm
미리	mm
밀리메터	mm
킬로그램	kg

- 4) 숫자, 기호, 영문표기에 대하여 표준발음과 달리 발생된 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)/(유네스코) : '유네스코'를 '유네코'로 잘못 발성한 경우

(UNESCO)/(유 엔 이 에스 씨 오) : '유네스코'를 '유 엔 이 에스 씨 오'로 알파벳으로 읽은 경우 (한국어)

(UNESCO)/(U N E S C O) : '유네스코'를 '유 엔 이 에스 씨 오'로 알파벳으로 읽은 경우 (영어)
UNESCO // '유네스코'라고 통상의 방식대로 발성한 경우

[(다섯)/(다섯) 대] // '다섯 대'를 '다섯 대'로 잘못 발성한 경우

아. 띄어쓰기

- 1) 띄어쓰기는 표준어법에 맞추어 하되 표준어법으로 명확히 결정할 수 없는 경우에는 띄운다.
2) 한글의 경우 성과 이름은 붙인다. 영어 이름은 이름 따로 성 따로 띄운다.

예) 이순신, 빌 클린턴

자. 알아듣기 힘든 발음

- 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 발생된 경우는 ‘*’를 붙이지 않는다.

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

나는(이렇게)/(이럴꼬) 그것을 해결하였다. (청취시 ‘이럴꼬’가 분명히 들리는 경우)

2) 방언에 해당하는 발성은 다음과 같이 이중 전사를 한다.

예) (장익사)/(장으사), (학교)/(핵교)

3) 문맥을 고려해봐도 전혀 알아들을 수 없는 발화는 'unk/' 으로 표기한다.

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

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

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.

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

7) 대화체 문장은 문장 자체가 이상하더라도 그대로 전사한다.