GUIDELINES FOR SPEECH CORPUS CREATION

Recording, Transcription, Metadata

As Part of Consortium Project titled Speech Technologies in Indian Languages Headed by Prof. Hema A Murthy & Prof. S Umesh, IIT Madras

Speech Quality Control Team

KONERU LAKSHMAIAH EDUCATION FOUNDATION

Vaddeswaram, Andhra Pradesh, India

Table of Contents

SPEECH CORPUS

Scope of the Project – "Speech Quality Control"

The objective of this project is to ensure the quality assessment and assurance of Speech Corpus. This helps in building high quality speech systems for the. Indian languages which can be used in various applications. This involves verifying the metadata, recording of speech data and transcription.

List of Indian languages Under Consideration:

Hindi, Indian English, Tamil, Telugu, Bengali, Gujarati, Marathi, Assamese, Kannada, Malayalam, Odia, Punjabi, Bodo, Manipuri, Rajasthani, Urdu, Kashmiri, Konkani, Sanskrit, Santali, Maithili, Khasi.

File Structure and Naming Convention:

Metadata File Name: "languageCode_vendorCode_submissionDate_releaseBatch.xlsx"

example: ***od_VENCD_08012023_r001.xlsx***

Speech File Name: "languageCode_vendorCode_speechMode_releaseBatch_sequenceId.wav"

example: ***od_VENCD_C_r001_s001.wav***

Transcription File Name: "languageCode_vendorCode_speechMode_releaseBatch_sequenceId.json"

example: ***od_VENCD_C_r001_s001.json***

ConventionCode	Description
languageCode	Can be with field width of 2 alphabet character in lowercase w.r.t below stated languageCode table.
vendorCode	Can be with field width of 5 alphaNumeric characters (alphabets should be in uppercase w.r.t this field).
submissionDate	Should be in ` ddmmyyyy `.
speechMode	Can be ` R / E /C `.
releaseBatch	Can be with field width of 3 numeric characters appended to character `r`, i.e. r001 – r999. Each release can contain `n` number of wav files in a particular releaseBatch.
sequenceId	Can be with field width of 3 numeric characters appended to character `s`, i.e. s001 – s999. This states the order of speech or transcription file in a particular releaseBatch.

languageCode	languageName	languageCode	languageName
hi	Hindi	**od**	Odia
in	Indian English	**bo**	Bodo
ta	Tamil	**mn**	Manipuri
te	Telugu	**rj**	Rajasthani
be	Bengali	**ur**	Urdu
gu	Gujarati	**ks**	Kashmiri
mr	Marathi	**ko**	Konkani
as	Assamese	**sa**	Sanskrit
kn	Kannada	**sn**	Santali
ml	Malayalam	**mt**	Maithili
pa	Punjabi	**kh**	Khasi

Note: In Future further languages code may be added later.

speechMode (alphabet should be in caps w.r.t this field)	
mode	**code**
Read Speech Data	R
Extempore Speech Data	E
Conversational Speech Data	C

Data Delivery:

Data Collection and Transcription Workflow should be submitted to Speech Quality Control Team. Data Delivery can be through any means, but data shouldn't be compressed by any means. Before delivering the data, make sure that the data or technologies used shouldn't violates any copyright infringement from the content providers.

Quality Check and Feedback:

All recordings will undergo quality assessment and validation. Only validated metadata, speech recordings and transcription will be accepted. Standards for acceptance criteria such as WER and TER will be as suggested by IITM. The SQC team will check and provide detailed feedback in order to accept or reject a particular batch of recordings and transcription. Assessment also aids us in marking the profanities specific to that language or region by employing the profanity checker.

Copyright Infringement: Consent Form needs to be collected per Participating Speaker.

GUIDELINES FOR METADATA

The following are the required metadata attributes for the Speech Corpus. We recommend the corresponding team to take utmost care of the following fields and their corresponding values at time of creation. The Sample Metadata file for Monologue as well as Conversational Speech will be provided as an additional document.

<table>
<tr><th colspan="2">Metadata Attributes / Field Names and Description</th></tr>
<tr><th>Attribute</th><th>Description</th></tr>
<tr><td>SL No.</td><td>It should be an Numeric character starting with digit ` 1 ` for each and every recording.</td></tr>
<tr><td>speechFile</td><td>Naming convention should be followed which as mentioned earlier.</td></tr>
<tr><td>Transcription File Name</td><td>Naming convention should be followed which as mentioned earlier.</td></tr>
<tr><td>callDuration</td><td>Should follow the convention in ` seconds.milli_seconds `. Milli Seconds should be considered up to 3 decimal points.</td></tr>
<tr><td>speechMode</td><td>Mention whether it’s an ` R / E / C `.</td></tr>
<tr><td>Topic</td><td>Theme of the discussion / Speech of the speaker.</td></tr>
<tr><td>SL No. of Speaker</td><td>It should be an Numeric character starting with digit ` 1 ` for each set of speakers in a conversation and goes up to number of speakers in a particular recording. It should be again reset to ` 1 ` in the next recording.</td></tr>
<tr><td>speakerUniqueId</td><td>It should be a string and the string can contain either alpha or numeric or both and it should be Unique Speaker id across the system.</td></tr>
<tr><td>Name of the Speaker</td><td rowspan="13">Details mentioned as per Consent Form.

If Mother Tongue Code, doesn’t exist in above given languageCode table then you can write specific name of the Mother Tongue itself.</td></tr>
<tr><td>Age</td></tr>
<tr><td>gender</td></tr>
<tr><td>Qualification</td></tr>
<tr><td>Occupation</td></tr>
<tr><td>Mother Tongue Code</td></tr>
<tr><td>Native Place</td></tr>
<tr><td>Current Location</td></tr>
<tr><td>District</td></tr>
<tr><td>State</td></tr>
<tr><td>Dialect - Zone</td></tr>
<tr><td>Dialect</td></tr>
<tr><td>Recording Device Type</td><td>Ex. G-Meet, Telephonic Conversation, Smartphone with Call Recording Features, Feature Phone with Call Recording facility, or any other recording app.</td></tr>
<tr><td>Recording Environment</td><td>Describe the place in which the speech has be recorded. Ex. Railway Station, Clean Environment, Bus Stand, Beach, Airport, In Vehicle, Open Traffic, Public Area.</td></tr>
<tr><td>Channel Type</td><td>Should be Mono.</td></tr>
<tr><td>Sampling Frequency (Hz)</td><td>For Wideband (Non Telephony) it should be 16000, but for Narrowband (Telephony) it can be either 8000 or else 16000.</td></tr>
<tr><td>Bits/Sample</td><td>16</td></tr>
<tr><td>Network Type</td><td>2G/3G/4G/5G/Telephony</td></tr>
<tr><td>Signal Bandwidth</td><td>Depending upon the recoding type.</td></tr>
<tr><td>Remark’s</td><td>-*- If Any -*-</td></tr>
</table>

GUIDELIENES FOR SPEECH RECORDING

1. Instruction And Requirements

1.1. Diversity

- Gender: 50% male, 50% female, +/- 10%.
- At least 4 dialects per language.
- Age Groups: 18-60 years.
- Domains of audio: Weather, different types of news, entertainment, health, agriculture, education, jobs, BPO (Business Process Outsourcing) data can be a mockup or it could be actual data (if we can get this from any agencies) – also monolingual (at best a few words in English) – Aadhar, passport, ration card, services – banking, repairs.
- No Profanity words should be used in the recording.
- Data can be collected from anywhere (No Restriction).
- Speech content which is communally or politically or racially biased should be avoided.

1.2. Audio Properties

- **Audio format -** WAV format ('.wav').
- **Sampling Frequency -** For Wideband (Non Telephony) it should be 16000 Hz, but for Narrowband (Telephony) it can be either 8000 Hz or else 16000 Hz.
- **Audio Coding Scheme -** 16-bit Linear PCM.
- **Bits per Sample -** 16.
- **Channel Type-** Mono.
- Audio Should not be post processed or pre-processed (e.g. compression, reverb, normalization).
- Low Background Noise.
- No distortion.
- No audio clip off.
- All Audio files should be intelligible.
- Correct audio segmentation for transcription.
- Correct Transcription as per the audio.

1.3. Speech Mode

1.3.1. Monologue (Single Speaker Speech)

- Read Speech
- Extempore Speech

1.3.2. Conversational Speech

- 2 to 4 speakers chatting about a topic of interest (for example sports, news, weather, entertainment, politics, business, everyday problems like public transport, e-governance, government schemes like aadhar, dhan yojana, ayush bharat etc).
- Speakers should prepare a little before the conversation. Many times, in the whole conversation they keep saying: "you tell", "what's up", "okay" "what shall i say' and no real conversation. So please ask them to have some points in their mind before the start of the conversation.
- Each conversation could last about 3-5mins between different speakers.

- The number of males and females should be more or less equal in multiple sessions.
- The conversations can be made across various smartphones using voice/data calls.
- The conversation speech data should be both code mixed and code switched. Prefer if it is mix of English, and another local language. (e.g., Tamil+English (Tanglish) or Hindi+English (Hinglish) etc.)

NOTE:

- **For both Monologue and Conversational speech data identity of speakers must be changed but should be uniquely named throughout the project/system.**
- **It is expected to provide 2 – 3 seconds of silence before starting the actual speech.**
- **At least 10 different domains must be chosen.**
- **Min 10 mins of data per speaker and max 30 mins per speaker. In the overall corpus, no speaker should have more than 30 minutes of audio.**

PROPOSED VOLUME (in % of collected Speech)			
Monologue (Single Speaker Speech) (80%)		Conversational (2 - 4 Speakers) (20%)	
Narrowband (Telephony) (30%)	Wideband (Non Telephony) (50%)	Narrowband (Telephony) (0%)	Wideband (Non Telephony) (20%)

GUIDELINES FOR TRANSCRIPTION

1. Segmentation

Segmentation involves "timestamping" the audio file for each given speaker. Segments are important to indicate the structural boundaries of an audio file, such as the type of sound, turns, utterances, and phrases in a conversation. Segment boundaries also make transcription easier by allowing the transcriber to hear manageable chunks of speech.

1.1. General Segmentation Requirements

- Create segments of <=15 seconds each for individual files larger than 30 seconds·
- Create segments (i.e. timestamping an audio file) according to the sound types.
- Each segment will be timestamped to the milliseconds. Timestamps must be positive floating numbers, in the format of seconds.milliseconds up to three decimal points (e.g., 10.355 for 10 seconds and 355 milliseconds).
- Each segment should have only one primary sound type, which will be listed as the primaryType — one of the segment objects — in the transcription JSON.
- Create each segment for its targeted sound signal with a 200-400 milliseconds padding at the start and end of a sound signal. Leave out continuous stretches of silence/white noise that last two or more seconds at the beginning, in the middle, or at the end of the segment.
- Transcription is needed only for Speech segments.

1.2 Sound Types for Segments

Barring silence/white noise, an audio file typically consists of the following sound types:

- Speech - Intelligible speech from a human or media source, regardless of whether the speaker is identifiable or not.
- Babble - Create Babble segments for audio signals that consist of speech or isolated vocal noise (e.g. coughing, laughing) from one or more background speakers (e.g., people standing nearby or in the same room), even if the speech is partially intelligible. This requires no transcription and the segment must be left blank with segment type labelled as Babble.
- Music – Audio from singing or musical instruments (including theme songs or characters singing songs).
- Noise – Any nonSpeech noise, either from humans or machines.
- Overlap – Any overlap intelligible/unintelligible speech between two or more speakers need no transcription and the segment must be left blank with segment type labelled as overlap.

1.3 Segment Labelling

Each segment must contain a list of segment objects. Some objects must be present and filled regardless of the primary type of a segment. Other objects must be present and filled for Speech segments only and excluded from other segment types. This must be present in a corresponding **JSON** file for an audio file.

For all segments, the following objects must be present and filled:

- **Start time -** Start timestamp of the segment in the format of seconds.milliseconds.
- **End time -** End timestamp of the segment in the format of seconds.milliseconds.
- **End of a segment timestamp should be the start of next segment timestamp.**
- **Segment ID -** A string that uniquely identifies the segment. E.g. UUIDs.
- **Loudness level -** One of the three loudness levels: Loud, Normal, or Quiet. Use "Normal" if not known.
- **Primary Sound Type -** One of the five sound types: Speech, Babble, Music, Noise, Overlap.
- Additionally, for Speech segments only, the following objects must be present and filled
- **Language code -** The language code of the language spoken in the segment should be in two-letter as per the language table provided in **`File Structure and Naming Convention`** section.
- **Speaker ID -** A string that uniquely identifies the speaker. We use UUID. The Speaker ID must be consistent throughout the entire file.
- **Transcription -** Data Transcription of the speech signals.
- **Speaker Diaziration -** In conversational speech accurate results of Speaker Diaziration needs to be provided. In the transcription, the identity of the speaker in each and every segment that contains the speech of (not more than) one speaker needs to be provided.

2. Transcription Conventions

All spoken words, including hesitations, filler words, false starts, and other verbal tics, should be captured accurately in the transcription.
Unless otherwise provided, verbatim transcription is the default setting.

2.1. Characters and Special Symbols

Only capital and lowercase letters, apostrophes, commas, exclamation points, hyphens, periods, question marks, and a limited number of unique markup symbols should be used in transcription.

Don't use numerals (e.g., 1, IV) and special symbols (e.g., $, +, @) to transcribe spoken words.

- "I have like $0" = "I have like zero dollars."
- "It was great/weird" = "It was great slash weird."
- "6 + 6 = 12." = "six plus six equals twelve."
- "My email is i-ruby@gmail.com" = "My email is I dash ruby at gmail dot com."

Below is the set of special mark-up symbols used in the transcription to indicate certain features or events within an audio file (e.g., unintelligible speech, code-mixing). Do not use these symbols for any reason other than as mark-up language.

Symbol(s)	Name	Use
< >	Angle brackets	Around opening and closing tags e.g., <initial>.
:	Colon	In conjunction with angle brackets and slash for non-target language tag e.g., <lang:Foreign></lang:Foreign>.
(())	Double parentheses	Around unintelligible speech or word
#	Hash tag	In front of filler words (aka, filled pauses).
/	Slash	In conjunction with angle brackets for closing markup tags e.g., </initial>.
[ ]	Square brackets	Around nonSpeech tags such [music].
~	Tilde	To indicate truncated speech.

2.2. Spelling and Grammar

Use standard spelling instead of phonetic spelling to transcribe what the speaker is saying.

2.2.1. Dialectal Pronunciations

Transcribe dialect pronunciations using the spellings of the "standard" forms, unless these dialect pronunciations are codified in an accepted written version of the dialect.

- "Issall well n' good darlin'." = "It's all well and good darling."
- "I'm from the wes' side." = "I'm from the west side."

2.2.2. Mispronounced Words

Transcribe mispronunciations using the standard spelling.

- "Call your representive." = "Call your representative."

2.2.3. Non-Standard Usage

Verbal transcription of the speaker's speech, even in cases where the speaker's words do not conform to the standard grammar of the language. Do not correct grammatical "errors" or speaker-created variations.

- "He been done work." = "He been done work."
- "We be playing basketball after work." = "We be playing basketball after work."

Check the spelling of all transcription files once the transcription is complete. If in doubt about the spelling of a word or name, for English, consult the American Heritage Dictionary: https://ahdictionary.com/. To consult the titles of songs, movies, TV shows, brands, etc., use a trusted site on the Internet.

2.3 Capitalization

Transcriptions must follow accepted capitalization patterns. For example, capitalize the first word of a sentence, proper nouns, and brand names

- "I want to visit Oregon" = "I want to visit Oregon."
- "I work at NASA" = "I work at NASA."
- "I'm going to Mexico on Thursday" = "I'm going to Mexico on Thursday."

2.4. Abbreviations

Do not introduce abbreviations in the transcription. Always spell out the full word when pronounced as such.

- "He's 6 ft 2!" = "He's six foot two."
- "Talk to Doctor Smith immediately." = "Talk to Doctor Smith immediately."

Use an abbreviation only if the speaker explicitly pronounces the word as abbreviated. Don't add a period after an abbreviated word (unless it appears at the end of a sentence).

- "I live in Cambridge, Mass." = "I live in Cambridge, Mass."
- "Billie Jean King went to Cal State." = "Billie Jean King went to Cal State."

The titles Ms, Mrs, Mr, and Mx that prefix a person's name are considered words in their own right, not abbreviations. When used as titles, transcribe them as Ms, Mrs, Mr, and Mx.

- "Mr. Smith this way please." = "Mr. Smith, this way please."

When used as direct addresses (without a following name), transcribe them as spelled-out forms (e.g., mister or missus).

- "Hey mister can you help me with this survey?" = "Hey, mister, can you help me with thissurvey?"

2.5 Contractions

Standard contractions must be transcribed as they are pronounced (e.g., isn't, where's, y'all). Include the apostrophe in the spelling.

For example, transcribe the following contractions as a single word:

- gimme
- gonna
- gotta
- lemme
- wanna
- watcha
- kinda

2.6. Individual Spoken Letters

Transcribe individual spoken letters as capital letters, separated by a space.

- "My name is John – jay, oh, eich, en". = "My name is John J O H N."

This does not apply to initialisms (e.g., IBM, FBI).

2.7. Numbers

Spell out numbers in full, not with numerals, according to how the speaker says them. This applies to both cardinal (e.g., 0, 215) and ordinal numbers (e.g., 1st, 5th).

- "5" = "five"
- "5th" = "fifth"
- "11th precinct" = "eleventh precinct"
- "306" = "three hundred and six", "three oh six", "three naught six", or "three zero six", depending on how it was pronounced.
- "Play radio 109.4 FM" = "play radio one oh nine point four <initial>FM</initial>"
- "Beverly Hills, 90210" = "Beverly Hills nine oh two one oh"

When spelling out numbers, use hyphens as required by the rules of the language. In English, numbers from twenty-one through ninety-nine are spelled with hyphens. Others are not hyphenated.

- "twenty-five"
- "three hundred"
- "five hundred fifty-two"
- "nineteen forty-five"

2.8. Punctuation

Only apostrophes, commas, exclamation points, hyphens, periods, question marks should be used as punctuation marks. Don't use any other English punctuation (e.g., semi-colons, and quotation marks).

Use these punctuations as required by the grammar rules.

End Punctuations	
Periods	Use a period only at the end of a complete sentence that is a statement. • I will skip my coffee today.
Question marks	Use a question mark only after a direct question or a tag question. • Isn't that simple? • You know the answer, don't you?
Exclamation points	Use an exclamation point at the end of a sentence when you feel or hear an emphatic stress or intonation. An exclamation point usually marks an outcry or an emphatic or ironic comment. • That's the biggest pumpkin I have ever seen! • When will I ever learn!

Sentence-Internal Punctuation	
Commas	Use commas to break up long stretches of speech. This is to facilitate reader comprehension. Below are some suggestions of when a comma should be used: • To separate items in a list of three or more, using the serial (aka Oxford) comma(i.e., the comma before the conjunction that joins the last two elements: o I enjoy skydiving, snowboarding, and mountain biking. • To set off a direct address: o Maryam, listen to me carefully. o I'm not calling you, my friends, just to whine about my life. • To break up compound and complex sentences: o I would like to join you, but I'm afraid I have class at that time. o Marcos and I couldn't go to the jazz concert, so we watched it on TV instead. • To set off introductory words and phrases: o Therefore, they cancelled their trip. o After taking a break, the team resumed their meeting. • Around parenthetical phrases: o That report on the New York Times was, to say the least, a bombshell. o Getting an exciting deal, like the one last year, would be awesome.

Word-Internal Punctuations

Apostrophes	Use apostrophes in contractions, possessives of individual letters, possessive "s", or as part of a person's name. • "That's where it's at" = "That's where it's at." • "Project Q's timeline" = "Project Q's timeline." • "Sinead O'Connor" = "Sinead O'Connor." • "Eleven o'clock" = "Eleven o'clock." • "Read Jess' email" = "Read Jess' email."
Hyphens	Use hyphens according to standard orthographic rules of the language. If it is not clear if a compound word should be spelled with a hyphen or not, Reference the American Heritage Dictionary as a reference. Here are a few examples of English compound words that can (or sometimes must) use hyphens: • a-line • d-day • ex-boyfriend, ex-drummer • extra-loud • self-aware • t-shirt • u-turn • v-neck • x-ray For product names, only use hyphens if they are parts of the official product names. • "Let's go to Chick-fil-A" = "Let's go to Chik-fil-A."

When transcribing a language other than English, use punctuation symbols and rules that are appropriate for that language. This could happen when a speaker switches to a foreign language in the middle of a segment. In this case, the foreign punctuation symbols should be within the foreign language tags <lang:Foreign></lang:Foreign>

• Hey, y'all. <lang:Spanish>¡Hola! ¿Cómo estás?</lang:Spanish> Sorry I'm late.

Note: Some punctuation use is stylistic/subjective.
Differences of opinion are not necessarily errors.

2.9. Acronyms and Initialisms

Acronyms refer to terms based on the initial letters of their various elements and are spoken as words. They should be transcribed as words in upper case without white spaces or periods between the letters. Standard reference: https://ahdictionary.com/

• "I work for NASA." = "I work for NASA."
• "AIDS has a great impact on society." = "AIDS has a great impact on society."
• "COVID has become a global pandemic." = "COVID has become a global pandemic."

Initialisms refer to terms spoken as series of letters (e.g., IBM, IMDB, HTTP). Initialisms should be written as upper case letters enclosed within the <initial> and </initial> tags.

- "I work for IBM." = "I work for <initial>IBM</initial>."
- "I like ZZ Top." = "I like <initial>ZZ</initial> Top."
- "http://www.gmail.com/" = "<initial>HTTP</initial> colon slash slash <initial>WWW</initial>dot gmail dot com."

Use periods only for initials standing for given names (e.g., E. B. White, George W. Bush). Otherwise, no period is needed in initialisms.

- "George W Bush paints now" = "George <initial>W.</initial> Bush paints now."

Do not include plural markers (e.g., -s) or the possessive marker ('s) within the <initial></initial> tags.

- "Welcome to the Ordinary Wizarding Level Examinations. O. W. L.s. More commonly known asOwls." = "Welcome to the Ordinary Wizarding Level Examinations. <initial>OWL</initial>s. More commonly known as Owls."

- "George W's dog was a Scottish Terrier." = "George <initial>W.</initial>'s dog was a ScottishTerrier."

Initialisms are treated as words. So, don't break up an initialism with any tags and don't include any other tags within the <initial></initial> tags.

- "I'll be taking my S (cough) AT next month." = "I'll be taking my [cough] <initial>SAT</initial>next month."

Notes:

- The word "OK"/"okay" is always transcribed as "okay. "
- Spoken individual letters (e.g., **proper names that are spelled out**) **are not initialisms** and don't require the <initial></initial> tags.

2.10. Disfluent Speech

Disfluent speech refers to any interruption of the normal flow of speech. Speakers may stumble over their words, repeat themselves, utter truncated words, restart phrases or sentences, and use hesitation sounds (i.e. filler words).

2.10.1. Stumbled Speech, Repetitions, and Truncated Words

Make your best effort to transcribe stumbled speech and repetitions according to what you hear after listening to the segment a few times.

- "Directions to the… to the… the hotel" = "Directions to the to the the hotel."

Use tildes (~) to indicate truncated words, whether at the beginning or the end.

- "Ale… alexa … stop the mu… the music." = "Ale~ Alexa, stop the mu~ the music."
- "...lexa play Janet Jackson… no wait…" = "~lexa, play Janet Jackson. No, wait."
- "N… n… no. It's Ch… Chom… Chomsky who said that." = "N~ n~ no. It's Ch~ Chom~ Chomsky who said that."

2.10.2. Filler Words

Filler words are "words" that speakers use to indicate hesitation or fill a pause in order to maintain control of a conversation while thinking of what to say next.

Each language has a limited set of filler words that speakers can use. For example:

EN_In	Gujarati	Marathi
#ah	#આહ	#आह
#er	#અર	#अर
#hm	#હમ	#हम
#uh	#ઉહ	#अह
#um	#અમ	#अं

Filler words may vary depending on the language being used in the speech content.

2.11. Unintelligible Speech

Use double parentheses (()) to mark stretches of speech that is difficult or impossible to understand or transcribe (such as when a speaker is speaking too softly or when there is noise over the speech). There should be a space before and after the double parentheses, but not within the parentheses themselves.

- "Alexa play ???? on spotify." = "Alexa, play (()) on Spotify."

If the transcriptionist has a guess about the speaker's words, transcribe what they think they hear within the double parentheses.

- "Alexa read ????? from audible." = "Alexa, read ((Cat In The Hat)) from Audible."
- "Alexa turn the ????" = "Alexa, turn the ((lights off))."

2.12. Non-Target Languages

When a speaker speaks Profanity word(s) specific to a language or region, place the tag <word:Profanity> at the location when the switch between languages begins and </word:Profanity> when the switch ends. When a segment contains the opening <word:Profanity> tag, it must also contain the closing </word:Profanity > tag.

When a speaker switches to a language other than English, place the tag <lang:Foreign> at the location when the switch between languages begins and </lang:Foreign> when the switch ends. When a segment contains the opening <lang:Foreign> tag, it must also contain the closing </lang:Foreign> tag.

If the transcriptionist can unambiguously identify the non-target language, replace "Foreign" with the language name in the tags. Capitalize the first letter of the language name.
Transcribe the speech of the non-target language, using the standard orthography of the nontarget language, if the transcriptionist understands the language. Otherwise, transcribe the nontarget language as (()).

- "You have to finish todo esto, porque. I have other things to do." = "You have to finish <lang:Spanish>todo esto, porque</lang:Spanish>. I have other things to do."

• "I'd like to tell her que ya no la quiero." = "I'd like to tell her <lang:Foreign>(())</lang:Foreign>."

Words of non-target language origin adopted into common use in the target language (i.e. loanwords) should be transcribed using the standard orthography of the target language. Don't use the <lang:Foreign></lang:Foreign> tags around loanwords that have been grammaticalized and fully adopted into common use in English. If it is unclear whether a word is a loanword or not, consult a dictionary like the American Heritage Dictionary: https://www.ahdictionary.com/. A word that is listed in the dictionary is a strong ground to consider it an established loanword, even if it is of foreign origin.

• "There was a tsunami in Indonesia." = "There was a tsunami in Indonesia."
• "Alexa… recipe for tacos" = "Alexa, recipe for tacos."
• "Remind me to spritz the flowers at eight." = "Remind me to spritz the flowers at eight."

Do not break up a word with the foreign language tags. This is rare in English, but in cases where a speaker mixes languages within a single word, such as having the root word in the non-target language but the affix in the target language:

1. Transcribe the word as it was pronounced using the respective standard orthography of each language.
2. Enclose both the root and the affix within the <lang:Foreign></lang:Foreign> tags.

2.13. NonSpeech

2.13.1. NonSpeech Noises

Indicate the following nonSpeech noises in the transcription by inserting the following tags in square brackets in the location where it occurs.

Tags	Category	Descriptions
[cry]	Human vocal noise	Crying/sobbing
[laugh]	Human vocal noise	Laughing, chuckling
[cough]	Human vocal noise	Coughing
[applause]	nonSpeech	Clapping
[nonHumanVocal]	nonSpeech/ nonHuman Sound	It can be dog bark's, or some animal sounds.
[silence]	Other noise type	Silence
[bgSpeech]	Other noise type	Speech in the background while speakers are not speaking
[music]	Other noise type	Music that is one or more seconds long without anyone speaking in the foreground. This includes on-hold music, songs, or singing

		Note: Don't use this tag for music playing in the background while someone's speaking
[noise]	Other noise type	Other miscellaneous noises not covered on the list above (e.g., screaming, coughing, sneezing, lipsmack, raining, punching, mic scratching, machine noise, telephone keypad, static noise, telephone ringing, etc.)
[overlap]	Other noise type	When more than one speaker speaks simultaneously at the same time.

Don't insert a nonSpeech tag in the middle of a word. If a nonSpeech sound occurs in the middle of a word, add the tag exactly before the word in which it occurred.

• "I will abso-(noise)-lutely open it" = "I will [noise] absolutely open it."

If a nonSpeech sound occurs repeatedly, represent it only once.

• "Wait … click click click click there" = "Wait [noise] there."

2.13.2. Silence/Pauses

Use the [no-speech] tag to indicate pauses or silence

• "They're not (pause) coming." = "They're not [no-speech] coming."

UNDERSTANDING JSON FORMAT

1. JSON Snapshot for Transcription of Monologue Speech Wave File

```
[
 {
    "speechMode": "R", // this attribute can be readspeech or extempore or conversational, and should follow the convention mentioned earlier for Speech Mode
    "recordingId": "13", // optional and it can be decided by the vendor
    "speechFile": "mr_VENCD_R_r001_s001.wav", // should follow the convention as per `File Structure and Naming Convention` section
    "callDuaration": 249.121, // it should be in seconds.milli seconds
    "speakers": [
      {
        "gender": "M", // it should be in uppercase
        "speakerSeqId": "1",
        "speakerUniqueId": "6" // It should be a numeric and should contain a unique id
      }
    ],
    "segments": [
      {
        "segmentType": "speech",
        "segmentId": 11001, // segment id should maintain the sequence down the file
        "start": 2.784, // milli seconds should be consider at most 3 decimal points
        "end": 13.550,
        "speakerSeqId": 1,
        "transcription": "आ आ जसकी आपल्या देशामधे जस दोन भाग आहे एक तर ग्रमीण शिक्षण वेवस्था आनी दूसरी शहरी आ शिक्षण वेवस्था"
      },
      {
        "segmentType": "speech",
        "segmentId": 11002,
        "start": 13.550, // should be end timestamp of previous segment
        "end": 27.082,
        "speakerSeqId": 1,
        "transcription": "भले देश एकही आहे तारी पण आपण जार बघितला तर आपल्या देशा मधे आपलयलाल दोन्हीं ठिकाना वरच्या जे शिक्षण वेवस्था भरपूर आपल्याला वेगळा दिसतात"
      },
      {
        "segmentType": "speech", // it can be speech, nonSpeech
        "segmentId": 11003,
        "start": 27.082,
        "end": 38.315,
        "speakerSeqId": 1,
        "transcription": "जसं की आपण शहरांमध्ये पाहिलं तर आपल्याला जाण्या-येण्यामध्ये साठी किंवा जे शिक्षक आहेत जे चांगले दर्जाचे आ"
      },
      {
```

```
        "segmentType": "speech",
        "segmentId": 11004,
        "start": 38.315,
        "end": 43.917,
        "speakerSeqId": 1,
        "transcription": "आ आ शिकलेले किंवा चांगले चांगले त्यांच्याकडे <lang:Foreign>
certificate </lang:Foreign> असलेले शिक्षक आहेत"
      },
      .
      .
      .
      .
      .
      {
        "segmentType": "speech",
        "segmentId": 11018,
        "start": 136.215,
        "end": 147.111,
        "speakerSeqId": 1,
        "transcription": "तर आपल्याला हे करण्यासाठी ग्रामीण ठिकाणा शिक्षण चांगलं करण्यासाठी
[cough] आपल्याला भरपूर गोष्टी करू शकतात सरकारने बरच काही नवीन नवीन <lang:Foreign> policy
</lang:Foreign>"
      },
      .
      .
      .
      .
      {
        "segmentType": "nonSpeech",
        "segmentId": 11039,
        "start": 256.842,
        "end": 266.86,
        "speakerSeqId": 1,
        "transcription": "(())"
      },
      .
      .
      .
      .
      .
      {
        "segmentType": "speech",
        "segmentId": 11046,
        "start": 322.632,
        "end": 330.295,
        "speakerSeqId": 1,
        "transcription": "त्या विद्येचा अनुभव तो घेऊ शकतो असं बरंच काही गोष्टी आपण तो करू
शकतो"
      }
    ]
  }
]
```

2. JSON Snapshot for Transcription of Conversational Speech Wave File

```
[
  {
    "speechMode": "C",
    "recordingId": "15",
    "speechFile": "mr_VENCD_C_r001_s002.wav",
    "callDuaration": 251.345,
    "speakers": [
      {
        "gender": "F",
        "speakerSeqId": "1",
        "speakerUniqueId": "7"
      },
      {
        "gender": "M",
        "speakerSeqId": "2",
        "speakerUniqueId": "6"
      }
    ],
    "segments": [
      {
        "segmentType": "speech",
        "segmentId": 11000,
        "start": 0.000,
        "end": 6.054,
        "speakerSeqId": 2,
        "transcription": "हा हा जो <lang:Foreign> Division of labour
</lang:Foreign> जो"
      },
      {
        "segmentType": "speech",
        "segmentId": 11001,
        "start": 6.054, // should be end timestamp of previous segment
        "end": 7.275,
        "speakerSeqId": 2,
        "transcription": "काल मार्क्सचा काय ये तर तुला काय वाट आणि काय पडलीये त्याच्याविषयी"
      },
      .
      .
      .
      .
      {
        "segmentType": "speech",
        "segmentId": 11023,
        "start": 28.146,
        "end": 32.125,
        "speakerSeqId": 1,
        "transcription": "नाही <lang:Foreign> feel </lang:Foreign> पाहिजे बाबा एकच
काम दिले आणि त्या कामांमध्ये गुंतून राहिलाय"
      },
      {
        "segmentType": "nonSpeech",
        "segmentId": 11024,
        "start": 32.125,
```

```
        "end": 32.927,
        "speakerSeqId": 2,
        "transcription": "[overlap] आणि मग"
      },
      {
        "segmentType": "speech",
        "segmentId": 11025,
        "start": 32.927,
        "end": 35.874,
        "speakerSeqId": 1,
        "transcription": "दुसऱ्या याच्यामध्ये <lang:Foreign> intetest </lang:Foreign>
किंवा <lang:Foreign> skill </lang:Foreign> असतील"
      },
      .
      .
      .
      .
      {
        "segmentType": "speech",
        "segmentId": 11036,
        "start": 51.8333,
        "end": 57.242,
        "speakerSeqId": 1,
        "transcription": "हा मग त्याच्यावरून हा त्याच्यावरून त्यांच्या कामावरून जात नाव बंद पडलं
की"
      },
      .
      .
      .
      .
      {
        "segmentType": "speech",
        "segmentId": 11061,
        "start": 98.3418,
        "end": 100.5,
        "speakerSeqId": 2,
        "transcription": "जास्त म्हणजे माणूस सोडून तो एक"
      },
      .
      .
      .
      .
      {
        "segmentType": "nonSpeech", // nonSpeech which means the utterance
which can't be transcribed
        "segmentId": 11158,
        "start": 335.841,
        "end": 336.057,
        "speakerSeqId": 0, // id should be zero
        "transcription": "[noise]"
      },
      .
      .
      .
      .
```

```
      {
        "segmentType": "speech",
        "segmentId": 11175,
        "start": 388.062,
        "end": 396.377,
        "speakerSeqId": 2,
        "transcription": "म्हणजे एक एक एक नाना आहे त्याची दोन बाजू आहे जर तुम्ही हे बघितलं तुम्हाला असं वाटणार तर तो बघितलं तर तसं वाटणार"
      },
      {
        "segmentType": "speech",
        "segmentId": 11176,
        "start": 396.979,
        "end": 403.386,
        "speakerSeqId": 2,
        "transcription": "[overlap] अच्छा ठीक आहे चालेल अच्छा ठीक आहे"
      }
    ]
  }
]
```


--*--THE END--*--