[**Some orienting notes** 4](#_Toc202285093)

[**Metadatafiles** 4](#_Toc202285094)

[ItsFileDetailsWOldFN.csv 4](#_Toc202285095)

[ItsFileDetailsShareable.csv 4](#_Toc202285096)

[MetadataInfAgeAndID.csv 5](#_Toc202285097)

[MergedTSAcousticsMetadata.csv 5](#_Toc202285098)

[FNSTETSimplified.csv 5](#_Toc202285099)

[FilesWithUnannotatedSections.csv 5](#_Toc202285100)

[FilesWithLessThan30minBnSections.csv 5](#_Toc202285101)

[ValDataMergedTSMetaDataTab.csv 5](#_Toc202285102)

[NumFilesSummary.csv 6](#_Toc202285103)

[TotAudioLengthAndKeySegPropDur.csv 6](#_Toc202285104)

[OlpSummaryNumbers.csv 6](#_Toc202285105)

[TotNumMerges\_NVocsTo1Voc\_ByVocType\_ReqAgesOnly.csv 6](#_Toc202285106)

[ZeroIviMergeDetailsTab\_<DataType>.csv 6](#_Toc202285107)

[**Sub-folders within the LENA daylong data folder (LENAData; contains LENA daylong data) and their contents** 7](#_Toc202285108)

[A1\_ItsFiles 7](#_Toc202285109)

[A2\_Segments 7](#_Toc202285110)

[A3\_PauseTimes 7](#_Toc202285111)

[A4\_TimeSeries 7](#_Toc202285112)

[A5\_TimeSeriesWPauses 7](#_Toc202285113)

[A6\_AcousticsTSJoinedwPauses 7](#_Toc202285114)

[A7\_ZscoredTSAcousticsLENA 8](#_Toc202285115)

[A8\_NoAcoustics\_0IviMerged\_LENA 8](#_Toc202285116)

[**Sub-folders within the human listener labelled data folder (HUMLabelData) and their contents** 8](#_Toc202285117)

[**A1\_HUMLabelData\_CleanupPipeline** 8](#_Toc202285118)

[**A2\_HUMLabelData\_PostCleanUp** 9](#_Toc202285119)

[A1\_CleanedUpEafFiles 9](#_Toc202285120)

[A2\_HlabelCsvFiles 9](#_Toc202285121)

[A3\_HlabelsOlpProcessed 9](#_Toc202285122)

[A4\_HlabelTS\_OlpProcessed 9](#_Toc202285123)

[A5\_HlabelTS\_OlpRemoved 10](#_Toc202285124)

[A6\_TSwSubrecsMerged 10](#_Toc202285125)

[A7\_HlabelTS\_Zscored 10](#_Toc202285126)

[A8\_MatchedLENAZscoreSections 10](#_Toc202285127)

[A9\_NoAcoustics\_0IviMerged\_Hum 11](#_Toc202285128)

[A10\_NoAcoustics\_0IviMerged\_L5min 11](#_Toc202285129)

[A11\_NoAcoustics\_0IviMerged\_Hum\_ChildDirANOnly 11](#_Toc202285130)

[**Files and sub-folders within the results folder (ResultsTabs; contains analyses results, reliability results, and results from data summary analyses)** 11](#_Toc202285131)

[**ResponseAnalyses** 12](#_Toc202285132)

[ResponseEffect\_w\_CurrPrevStSizeControl\_<DataType>.csv 12](#_Toc202285133)

[RespEff\_NoPrevStSizCtrl\_VarsScaleLog\_CorpusLvl\_IviOnly\_CI99\_9prc.csv, RespEff\_W\_PrevStSizCtrl\_VarsScaleLog\_CorpusLvl\_IviOnly\_CI99\_9prc.csv 12](#_Toc202285134)

[PrevStSizeResids\_VarsScaleLog\_RecDayLvl\_IviOnly.csv 12](#_Toc202285135)

[RecLvlPrevStSizeAndRespBetas.csv 12](#_Toc202285136)

[AgeEffects\_IviOnly\_CI99\_9prc.csv 13](#_Toc202285137)

[AgeEffects\_IviOnly\_CI99\_9prc.xlsx 13](#_Toc202285138)

[TransformedIVIsAndResidsFromPrevIVILmer\_AgeBlockLvl\_LENA.csv 13](#_Toc202285139)

[AgeLvlPrevIVIBetaAndIntercept\_LENA.csv 13](#_Toc202285140)

[**DataDescriptionSummaries** 13](#_Toc202285141)

[NumFilesSummary.csv 13](#_Toc202285142)

[LdaySegmentLevelDurationSummaryStats\_No0IviMerge.csv 13](#_Toc202285143)

[LdayRecDayLevelSegmentNumsSummaryStats\_No0IviMerge.csv 14](#_Toc202285144)

[LdaySegsDurStruct\_No0IviMerged.mat 14](#_Toc202285145)

[LdaySegmentLevelDurationSummaryStats\_0IviMerged\_ChnAdOnly.csv, LdayRecDayLevelSegmentNumsSummaryStats\_0IviMerged\_ChnAdOnly.csv, LdaySegsDurStruct\_0IviMerged\_ChnAdOnly.mat 14](#_Toc202285146)

[LdaySegDurAndNumsSummaryStats\_OnlyChnAd0IviMerged.xlsx 15](#_Toc202285147)

[<ValidationDataType>SegmentLevelDurationSummaryStats\_0IviMerged\_ChnAdOnly.csv 15](#_Toc202285148)

[<ValidationDataType>RecDayLevelSegmentNumsSummaryStats\_0IviMerged\_ChnAdOnly.csv 15](#_Toc202285149)

[<ValidationDataType>SegsDurStruct\_0IviMerged\_ChnAdOnly.mat 15](#_Toc202285150)

[TotVoc<Dur/Nums>Summaries\_0IviMerged\_ChnAdOnly\_<DataType>.xlsx 16](#_Toc202285151)

[TotAudioLengthAndKeySegPropDur.csv 17](#_Toc202285152)

[TotNumMerges\_NVocsTo1Voc\_ByVocType\_ReqAgesOnly.csv 17](#_Toc202285153)

[**ReliabilityTabs** 17](#_Toc202285154)

[ReliabilityErrorRates\_FileOrSectionLvl.csv 17](#_Toc202285155)

[ConfusionMatStruct.mat 18](#_Toc202285156)

[ReliabilityAggregateTab.xlsx 18](#_Toc202285157)

[**UttDurTabs** 18](#_Toc202285158)

**Some orienting notes**:

Below, FileNameRoot\_subrec indicates that the relevant files are still at the sub-recording level, while FileNameRoot indicates that the relevant files are presented after sub-recordings have been merged, (eg. 0056\_000305a and 0056\_000305b merged into 0056\_000305).

In the code, file names, and in column names within data files, both infants and adult caregivers are referred to as speakers. This is a broad descriptor for vocalisations rather than speech and does not imply that infants as young as 3 months old have developed speech.

Below, the terms IEI and IVI are used interchangeably to mean inter-event intervals (as described in the Burstiness paper and pre-print) and equivalently, inter-vocalisation intervals, in that the relevant events are vocalisations.

Similarly, the Burstiness paper and pre-print use the short-hands ChSp (infant speech-related), ChNsp (infant non-speech-realated), and Ad (adult) for vocalisation types analysed in or relevant to the study. Throughout the code, however, these vocalisations are indicated using the short-hands CHNSP (infant speech-related), CHNNSP (infant non-speech-realated), and AN (adult). In addition, CHN indicates the combined CHNSP (infant speech-related) and CHNNSP (infant non-speech-realated) category.

For a summary of data organization and info about which scripts output which files, see Metadata\_CodeAndFiles.xlsx. Some files are not shared because they contain potentially identifying info. For more info about which files are not shared, see Metadata\_CodeAndFiles.xlsx.

Finally, some of the results files are presented as .csv and .xlsx files. Both contain the same information but the .xlsx files have been manually re-organised to aid interpretation.

**Metadatafiles**: contains metadata files describing various aspects of the datasets analysed.

ItsFileDetailsWOldFN.csv: contains info about all .its files in the LENA daylong corpus (including some files at ages other than 3, 6, 9, and 18 months. These files are not included in any analyses presented in the Burstiness paper/pre-print), prior to sub-recordings being merged. As such, filenames contain sub-recording suffixes (see Section **Some Orienting Notes** in this file). Also contains filenames prior to de-identification and hence, is not shared.

*Colnames*: FNRoot (de-identified file name root), OldFNRoot (old file name root), InfantID

ItsFileDetailsShareable.csv: this is a version of ItsFileDetailsWOldFN.csv with the file names prior to de-identification removed.

*Colnames*: FNRoot (de-identified file name root), InfantID

MetadataInfAgeAndID.csv: contains info about all .its files in the LENA daylong corpus (including some files at ages other than 3, 6, 9, and 18 months. These files are not included in any analyses presented in the Burstiness paper/pre-print), prior to sub-recordings being merged. As such, filenames contain sub-recording suffixes (see Section **Some Orienting Notes** in this file).

*Colnames*: FNRoot (de-identified file name root), InfantID, InfantAge (in days)

MergedTSAcousticsMetadata.csv: contains info about all files in the LENA daylong corpus (including some files at ages other than 3, 6, 9, and 18 months. These files are not included in any analyses presented in the Burstiness paper/pre-print), after sub-recordings are merged. As such, filenames do not contain sub-recording suffixes (see Section **Some Orienting Notes** in this file). **This is the final, consolidated metadata file that describes the LENA daylong data analysed in the Burstiness paper/pre-print** (after filtering out files at ages other than 3, 6, 9, and 18 months).

*Colnames*: FileNameRoot (de-identified file name root after sub-recordings have been merged), InfantAgeDays, InfantAgeMonth, InfantID

FNSTETSimplified.csv: contains info about 5-minute sections tagged for annotation. This file was generated by hand by JM, who is a co-author on the Burstiness paper/pre-print and therefore, no code to generate this file is provided.

*Colnames*: FileName (the file the section belongs to, with sub-recording suffixes; see Section **Some Orienting Notes** in this file), StartTime, EndTime (start and end times of the 5-minute section to be annotated in hh:mm:ss), StartTimeSS, EndTimeSS (start and end times of the 5-minute section to be annotated in seconds, with the start of the relevant daylong recording being 0 s)

FilesWithUnannotatedSections.csv: lists files that have at least one unannotated 5-minute section (which was tagged to be annotated in FNSTETSimplified.csv)

*Colnames*: Fname\_UnannotSections (file name with the unannotated section(s), after sub-recordings have been merged), NumSectionsToAnnot\_CodingSheet, NumSectionsAnnotated

FilesWithLessThan30minBnSections.csv: lists files that have at least one pair of 5-minute sections that are less than 30 minutes apart.

*Colnames*: Fname\_TimeBnSectionsFlag (relevant file name, after sub-recordings have been merged), Num\_TimeBnSectionsLessThan30min (number of instances when two 5-minute sections in the file are less than 30 minute apart)

ValDataMergedTSMetaDataTab.csv: contains info about all files in the validation dataset, after sub-recordings are merged. As such, filenames do not contain sub-recording suffixes (see Section **Some Orienting Notes** in this file). **This is the final, consolidated metadata file that describes the validation dataset(s) analysed in the Burstiness paper/pre-print**.

*Colnames*: FileNameRoot (de-identified file name root after sub-recordings have been merged), InfantAgeDays, InfantAgeMonth, InfantID, NumSections (number of annotated 5-minute sections in the validation data file)

NumFilesSummary.csv: summarises the total number of daylong files (LENA daylong data) and validation data files (validation datasets) in the relevant corpora after sub-recordings have been merged. This table is reproduced as Supplementary Table S1 in the Burstiness paper/pre-print.

*Colnames*: AgeMonth (infant age + descriptors for a couple extra rows), NumFiles\_Lday, NumFilesValData

TotAudioLengthAndKeySegPropDur.csv: contains info presented in Section S3.1 in the supplementary info for the Burstiness paper/pre-print and summarises various total durations (and relevant proportions) across the entire relevant dataset.

*Colnames*: DataType, TotAudioDur\_s (total audio duration in hours: total length of the daylong recording from recording start to recording end minus recorder pauses or deletions for the LENA daylong data; total duration of 5-minute sections annotated based on the section start and end time in FNSTETSimplified.csv for validation data), TotAudioDur\_hr (total audio duration in hours, converted from ), TotKeySegDur\_s (total duration of key segments—infant speech-related, infant non-speech related, and adult—segments combined, in seconds), TotKeySegDur\_hr, PercKeySegDurInTotAudio (proportion of total key segment duration wrt total audio duration)

OlpSummaryNumbers.csv: summarises overlap processing details for infant speech-related, infant non-speech related, and adult vocalisations for the human listener-labelled data (reproduced as Supplementary Table S5 in the Burstiness paper/pre-print).

*Colnames*: see Supplementary Table S5 in the Burstiness paper/pre-print

TotNumMerges\_NVocsTo1Voc\_ByVocType\_ReqAgesOnly.csv: summarises details of 0 s IEI merging for all datasets (reproduced as Supplementary Table S6 in the Burstiness paper/pre-print).

*Colnames*: see Supplementary Table S6 in the Burstiness paper/pre-print

ZeroIviMergeDetailsTab\_<DataType>.csv: contains more granular (file-level, after sub-recordings have been merged) details about 0 s IEI merging for the different datasets.

*Colnames*: InfantID, InfantAgeMonth, TotalMergeCt\_CHNNSP (number of merge events consisting of two CHNNSP vocalisations separated by a 0 s IEI being merged into a single vocalisation. If this is part of a merge chain, where *n* > 2 vocalisations are merged into one, each pair of vocalisations being merged counts as one merge event), TotalMergeCt\_CHNSP, TotalMergeCt\_AN

**Sub-folders within the LENA daylong data folder (LENAData; contains LENA daylong data) and their contents:**

(Note that the LENA daylong files contains 4 files at ages other than 3, 6, 9, and 18 months. These get processed as part of the pipeline but analyses are only carried out on daylong recordings at aged 3, 6, 9, and 18 months)

A1\_ItsFiles: contains LENA .its (interpreted time segments) files.

*File name format*: <FileNameRoot\_subrec>.its

A2\_Segments: contains segments files with start (startsec) and end (endsec) times in seconds as well as sound type labels (segtype) of all sound segments identified by LENA.

*File name format*: <FileNameRoot\_subrec>\_Segments.csv

*Colnames*: startsec, endsec, segtype

A3\_PauseTimes: contains .txt files with info about recorder pauses

*File name format*: <FileNameRoot\_subrec>\_PauseTimes.txt

*Colnames*: There are no column names in the file. However, the first column identifies the start time (in seconds elapsed in the audio file) and the end time of each subrecording within a .wav file.

A4\_TimeSeries: contains .csv files with acoustics (mean pitch: meanf0; mean amplitude: dB), duration (seconds), start and end times (seconds) of all infant (CHNSP, CHNNSP) and adult (AN) vocalisations in the file as identified by LENA. Also contains relevant speaker labels and identifiers for each audio segment corresponding to the vocalisations. Durations in this file for some audio segments may be NaN. This is because these correspond to audio segments for which the relevant script is unable to compute a mean pitch and/or amplitude. As such, durations are later recomputed based on vocalisation onset and offset.

*File name format*: <FileNameRoot\_subrec>\_TS.csv

*Colnames*: wavfile (audio segment identifier), speaker, start, end, duration, meanf0, dB

A5\_TimeSeriesWPauses: contains .csv files with from A4\_TimeSeries with an extra column indicating whether a vocalisation is the last one in a subrecording or not (SubrecEnd, 1 indicates that the vocalisation is the last one in a subrecording, 0 otherwise)

*File name format*: <FileNameRoot\_subrec>\_TSwPauses.csv

*Colnames*: wavfile, speaker, start, xEnd, duration, meanf0, dB, SubrecEnd

A6\_AcousticsTSJoinedwPauses: contains .csv files where subrecordings that were separated into separate .csv files but are from the same daylong recording are merged into a single file. These files have a column identifying the name of the file before this merging happens (FileNameUnMerged). Durations in these files have been re-computed based on start and end times of each vocalisation (see description for A4\_TimeSeries above for details). The information in the SubrecEnd column in A5\_TimeSeriesWPauses is now encoded in the column SectionNum, which identifies the unique ‘section’ (used to mean sub-recording, resulting from recorder pauses or deletions) that each vocalisation belongs to. Inter-event intervals (IEIs) or acoustic steps are not computed for vocalisations in different sections because there is no temporal continuity between different sections.

*File name format*: <FileNameRoot>\_AcousticsTSJoined.csv

*Colnames*: wavfile, speaker, start, xEnd, duration, meanf0, dB, SectionNum, FileNameUnMerged

A7\_ZscoredTSAcousticsLENA: contains .csv files with the duration, meanf0, dB columns from the files in A6\_AcousticsTSJoinedwPauses replaced by their z-scored values (z-scoring done across the entire LENA daylong and human listener-labelled datasets’ relevant acoustics combined). For duration and meanf0, the values have been log-ed (to the base 10) prior to z-scoring.

*File name format*: <FileNameRoot>\_ZscoredAcousticsTS\_LENA.csv

*Colnames*: wavfile, speaker, start, xEnd, logDur\_z, logf0\_z, dB\_z, SectionNum, FileNameUnMerged

A8\_NoAcoustics\_0IviMerged\_LENA: contains .csv files with only vocalisation onset and offset info (acoustics and duration values from A7\_ZscoredTSAcousticsLENA are removed at this stage) *after* vocalisations of the same label type (CHNSP, CHNNSP, or AN) separated by 0 s IEIs have been merged together. Also note that for these files, the FAN and MAN label categories have been merged into a single AN category (but for files before this point, the FAN and MAN label categories were identified separately).

*File name format*: <FileNameRoot>\_ NoAcoustics\_0IviMerged\_LENA.csv

*Colnames*: wavfile, speaker, start, xEnd, FileNameUnMerged, SectionNum

**Sub-folders within the human listener labelled data folder (HUMLabelData) and their contents:**

**A1\_HUMLabelData\_CleanupPipeline**: contains .eaf files and parsed .csv files before and after (automated and manual) clean-up rounds. Also contains summary data files for the .eaf files before and after clean-up rounds (see Metadata\_CodeAndFiles.xlsx for a summary of how the data is organized within this folder and which scripts output which files). No subfolders or files within this folder are shared.

.eaf files prior to some *n*-th clean-up round (which may include automated and manual clean-up) are in A1\_EAFFiles, parsed .csv files based on these .eaf files (prior to the current clean-up round) are in A2\_ParsedEafFilesFromR\_PreCleanUp, .eaf files after the current round of automated clean-up and associated editing are in A3\_EditedEafFiles, and parsed .csv files from those cleaned-up .eaf files are in A4\_ParsedEafFilesFromR\_PostCleanUp. Summary .csv and .txt files summarizing flagged annotations for various types of clean-up before and after the clean-up round are in SummaryCsvAndTxtFiles. SummaryCsvAndTxtFiles also contains the metadatafile (FNSTETSimplified.csv) that has info about 5-minute sections tagged for annotation, the file they belong to (FileName), and the start and end times of the 5-minute section to be annotated (StartTime, EndTime; in hh:mm:ss) and (StartTimeSS, EndTimeSS; in seconds, with the start of the relevant daylong recording being 0 s).

**A2\_HUMLabelData\_PostCleanUp**: contains .csv files parsed from .eaf files after clean-up rounds and subsequent files throughout the data processing pipeline.

A1\_CleanedUpEafFiles: final .eaf files after clean-up rounds, copied from the relevant subfolder within A1\_HUMLabelData\_CleanupPipeline (if no further manual clean-up rounds are performed on these files) or from the most up-to-date post-(automated and manual)-cleanup data folder.

A2\_HlabelCsvFiles: parsed .csv files from the cleaned up .eaf files.

*File name format*: <FileNameRoot\_subrec>.csv (but the suffix after the <FileNameRoot\_subrec> may change to identify different clean-up cycles)

*Colnames*: StartTimeRef (reference to vocalisation start time, eg. ts1, ts2, etc.), StartTimeLineNum (line number in the .eaf file where the start time is mentioned for the vocalisation in the annotation block for each vocalisation), EndTimeRef, EndTimeLineNum (similarly for end times), AnnotId (annotation id, eg. a1, a2, etc.), AnnotIdLineNum (line number with the annotation id in the annotation block for each vocalisation), Annotation (actual annotation), AnnotationLineNum (line number with the annotation in the annotation block for each vocalisation), TierTypeVec (the tier name for the vocalisation, eg. Infant Voc Type, Adult Utterance Direction, etc. Note that all tiers available in the .eaf files have been parsed into these .csv files), StartTimeVal, EndTimeVal (start and end times in ms for vocalisations as identified by human annotator)

A3\_HlabelsOlpProcessed: .csv files for files in with A2\_HlabelCsvFiles after overlapping vocalisations have been processed into non-overlapping sub-vocalisations and sub-vocalisations that are fully overlaps. The latter category (sub-vocalisations that are entirely overlaps) has been tagged accordingly (OLP in the Annotation column). These files contain only adult and infant vocalisations. Further, there is a vocalisation index (VocIndex) that identifies sub-vocalisations from the same original vocalisation by the index of the original (un-chopped) vocalisation.

*File name format*: <FileNameRoot\_subrec> \_OlpProc.csv

*Colnames*: StartTimeRef, StartTimeLineNum, EndTimeRef, EndTimeLineNum, AnnotId, AnnotIdLineNum, Annotation, AnnotationLineNum, TierTypeVec, StartTimeVal, EndTimeVal, VocIndex

A4\_HlabelTS\_OlpProcessed: contains .csv files with acoustics (mean pitch: meanf0; mean amplitude: dB), duration (seconds), start and end times (seconds) of all infant (CHNSP, CHNNSP) and adult (AN) vocalisations in the file as identified by the human annotator. Also contains relevant speaker labels and identifiers for each audio segment corresponding to the vocalisations. However, that annotation level labels are not present in these files and instead, the column only has details of whether the vocalisation is an adult (AN) vocalisation or the target infant (CHN, which encompasses CHNSP and CHNNSP). Annotation level labels as well as differentiated labels for CHNSP and CHNNSP are added back in at a later step. Durations in this file for some audio segments may be NaN. This is because these correspond to audio segments for which the relevant script is unable to compute a mean pitch and/or amplitude. As such, durations are later recomputed based on vocalisation onset and offset.

*File name format*: <FileNameRoot\_subrec>\_TSOlpProc.csv

*Colnames*: wavfile (audio segment identifier), speaker, start, end, duration, meanf0, dB

A5\_HlabelTS\_OlpRemoved: contains .csv files with acoustic time series and start and end times for all infant and adult vocalisations as identified by human listeners, after overlap processing and with all sub-vocalisations that are fully overlaps have been removed. Further, annotation labels have been added back and speaker labels have been recast as CHNSP, CHNNSP, and AN instead of only CHN and AN.

*File name format*: <FileNameRoot\_subrec>\_TS\_OlpRemoved.csv

*Colnames*: wavfile, speaker, start, xEnd, duration, meanf0, dB, Annotation

A6\_TSwSubrecsMerged: contains .csv files with acoustic time series from A5\_HlabelTS\_OlpRemoved *after* files from different subrecordings (but from the same daylong recording) have been merged. These files have a column identifying the name of the file before this merging happens (FileNameUnMerged) as well as a column identifying the unique 5-minute section the vocalisation belongs to (SectionNum). Inter-event intervals (IEIs) or acoustic steps are not computed for vocalisation in different sections because there is no temporal continuity between different sections. Durations in these files have been re-computed based on start and end times of each vocalisation (see description for A4\_HlabelTS\_OlpProcessed above for details).

*File name format*: <FileNameRoot>\_TSSubrecMerged.csv

*Colnames*: wavfile, speaker, start, xEnd, duration, meanf0, dB, Annotation, FileNameUnMerged, SectionNum

A7\_HlabelTS\_Zscored: contains .csv files with the duration, meanf0, dB columns from the files in A6\_TSwSubrecsMerged replaced by their z-scored values (z-scoring done across the entire LENA daylong and human listener-labelled datasets’ relevant acoustics combined). For duration and meanf0, the values have been log-ed (to the base 10) prior to z-scoring.

*File name format*: <FileNameRoot>\_ZscoredAcousticsTS\_Hum.csv

*Colnames*: wavfile, speaker, start, xEnd, logDur\_z, logf0\_z, dB\_z, Annotation, FileNameUnMerged, SectionNum

A8\_MatchedLENAZscoreSections: contains .csv files with LENA-labelled data corresponding to the human-listener labelled 5-minute sections. Here, the SectionNum column corresponds to the 5-minute annotated sections and matches the section numbers in the human-labelled data (A7\_HlabelTS\_Zscored). For e.g., section number 1 in the file corresponding to the daylong recording 009\_000302 in A7\_HlabelTS\_Zscored and A8\_MatchedLENAZscoreSections represent human and LENA labelled data for the same 5-minute section (and so on and so forth).

*File name format*: <FileNameRoot>\_MatchedLENA\_ZscoreTS.csv

*Colnames*: wavfile, speaker, start, xEnd, logDur\_z, logf0\_z, dB\_z, FileNameUnMerged, SectionNum

A9\_NoAcoustics\_0IviMerged\_Hum: contains .csv files with only vocalisation onset and offset info (acoustics and duration values from A7\_HlabelTS\_Zscored are removed at this stage) *after* vocalisations of the same label type (CHNSP, CHNNSP, or AN) separated by 0 s IEIs have been merged. Note that this merging does not treat different annotation labels for the same speaker label category differently. For e.g., if two CHNNSP vocalisations are separated by a 0 s IEI, with the first one being R and the second one being L, these two vocalisations are still merged into one and will retain the annotation label for the first vocalisation in the series (R, in this case). However, such cases are very rare and form a negligible fraction of the total number of vocalisations in each category. See the script A2\_1\_CtsOfMergedUnmatchedAnnots.m for details. Files in this folder contain all adult vocalisations as identified by human listeners (as opposed to files in A11\_NoAcoustics\_0IviMerged\_Hum\_ChildDirANOnly; see below).

*File name format*: <FileNameRoot>\_NoAcoustics\_0IviMerged\_Hum.csv

*Colnames*: wavfile, speaker, start, xEnd, Annotation, FileNameUnMerged, SectionNum

A10\_NoAcoustics\_0IviMerged\_L5min: contains .csv files with only vocalisation onset and offset info (acoustics and duration values from A8\_MatchedLENAZscoreSections are removed at this stage) *after* vocalisations of the same label type (CHNSP, CHNNSP, or AN) separated by 0 s IEIs have been merged.

*File name format*: <FileNameRoot>\_ NoAcoustics\_0IviMerged\_LENA5min.csv

*Colnames*: wavfile, speaker, start, xEnd, FileNameUnMerged, SectionNum

A11\_NoAcoustics\_0IviMerged\_Hum\_ChildDirANOnly: contains .csv files with only vocalisation onset and offset info (acoustics and duration values from A7\_HlabelTS\_Zscored are removed at this stage) *after* vocalisations of the same label type (CHNSP, CHNNSP, or AN) separated by 0 s IEIs have been merged. However, unlike files in A9\_NoAcoustics\_0IviMerged\_Hum, files in this folder exclude non-infant directed adult vocalisations (annotations U and N) as identified by human listeners. All other descriptions are the same as for files in A9\_NoAcoustics\_0IviMerged\_Hum.

*File name format*: <FileNameRoot>\_ NoAcoustics\_0IviMerged\_ChildDirANOnly\_Hum.csv

*Colnames*: wavfile, speaker, start, xEnd, Annotation, FileNameUnMerged, SectionNum

**Files and sub-folders within the results folder (ResultsTabs; contains analyses results, reliability results, and results from data summary analyses):**

**ResponseAnalyses**: contains .csv and .xlsx files with results from response analyses with and without previous IEI control and sub-folders containing .csv files with data files used in the response analyses. All analyses are only carried out for LENA daylong recordings at ages 3, 6, 9, and 18 months (the validation data only has files at ages 3, 6, 9, and 18 months).

ResponseEffect\_w\_CurrPrevStSizeControl\_<DataType>.csv: contains .csv files with current and previous IEIs, response data (for response windows from 0.5 s to 10 s), infant ID and infant age info for all datasets analysed in the Burstiness paper/pre-print, for both adult response to infants and infant response to adults.

*File name format*: CurrPrevStSize\_<DataType>\_<ResponseType>\_IviOnly.csv

*Colnames*: Response\_<ResponseWindow>, CurrIVI, PrevIVI, InfantID, AgeDays, AgeMonths

RespEff\_NoPrevStSizCtrl\_VarsScaleLog\_CorpusLvl\_IviOnly\_CI99\_9prc.csv, RespEff\_W\_PrevStSizCtrl\_VarsScaleLog\_CorpusLvl\_IviOnly\_CI99\_9prc.csv: response analyses results (with and without the previous IEI control) for all datasets analysed in the Burstiness paper/pre-print. The previous IEI beta columns are only present for the stats results table w/ the previous IEI control.

*Colnames*: RespWin\_Seconds, InfAge\_Months, PrevSt\_Beta, PrevStP, PrevStCI\_Lwr, PrevStCI\_Upper, Response\_Beta, ResponseP, ResponseCI\_Lwr, ResponseCI\_Upper, ResponseType, DataType

RespEff\_W\_PrevStSizCtrl\_VarsScaleLog\_CorpusLvl\_IviOnly\_CI99\_9prc.xlsx: Excel sheet copy of the corresponding .csv file with results manually sorted and highlighted at various significance levels for easier visual parsing. See the Notes sheet in the file for details.

PrevStSizeResids\_VarsScaleLog\_RecDayLvl\_IviOnly.csv: Contains IEI residuals as well as scaled and unscaled current and previous IEIs based on recoding day level (or validation data file level, for the validation data) previous IEI control analyses, for all data analysed in the Burstiness paper/preprint.

*Colnames*: Response\_<ResponseWindow>, CurrIVI, PrevIVI, InfantID, AgeMonths, CurrIVI\_Trans, PrevIVI\_Trans, ResidVar, ResidVar\_Scaled, ResponseType, DataType

RecLvlPrevStSizeAndRespBetas.csv: contains recording day level (or validation data file level, for the validation data) previous IEI betas, as well as response betas with and without the previous IEI control. This data is then used to compute the effect of infant age on these betas.

*Colnames*: RespWin\_Seconds, InfAge\_Months, InfID, PrevSt\_Beta, Response\_Beta\_wCtrl, Response\_Beta\_woCtrl, ResponseType, DataType

AgeEffects\_IviOnly\_CI99\_9prc.csv: contains results from performing age effects analyses on data in RecLvlPrevStSizeAndRespBetas.csv. Note that the reported effects are Bs (as opposed to betas, which are standardized regression coefficients).

*Colnames*: DataType, SpkrAndResp, RespWindow\_s, BetaType,

InterceptVal (intercept from the regression), InterceptCI\_Lwr, InterceptCI\_Upper, Age1Eff (linear age effect), Age1P, Age1CI\_Lwr, Age1CI\_Upper, Age2Eff (quadratic age effect), Age2P, Age2CI\_Lwr, Age2CI\_Upper

AgeEffects\_IviOnly\_CI99\_9prc.xlsx: Excel sheet copy of AgeEffects\_IviOnly\_CI99\_9prc.csv with results manually sorted and highlighted at various significance levels for easier visual parsing. See the Notes sheet in the file for details.

TransformedIVIsAndResidsFromPrevIVILmer\_AgeBlockLvl\_LENA.csv: Response data for the 5 s response window, scaled current and previous IEIs, and residual IEIs from previous IEI correlation estimation at the age block level (3, 6, 9, and 18 months) for the LENA daylong data

*Colnames*: Response\_5, CurrIVI, PrevIVI, InfantID, AgeMonths, ResponseType, ZscoreLog10\_CurrIVI, ZscoreLog10\_PrevIVI, ResidualVec

AgeLvlPrevIVIBetaAndIntercept\_LENA.csv: Intercepts and previous IEI betas for the data in TransformedIVIsAndResidsFromPrevIVILmer\_AgeBlockLvl\_LENA.csv at the age block level.

*Colnames*: InfAge, TypeOfResponse, Intercept, PrevStBeta

**DataDescriptionSummaries**: contains files (.csv, .mat, .xlsx) with results from data summary analyses scripts, summarising various aspects of the data. Summaries are reported only for LENA daylong recordings at ages 3, 6, 9, and 18 months unless otherwise specified (the validation data only has files at ages 3, 6, 9, and 18 months).

NumFilesSummary.csv: contains info about the number of files in the LENA daylong and validation datasets.

*Colnames*: AgeMonth, NumFiles\_Lday, NumFiles\_ValData

LdaySegmentLevelDurationSummaryStats\_No0IviMerge.csv: contains summary stats about the duration of each segment type (without merging segments of the same type separated by 0 s IEIs) in the LENA daylong corpus across ages (3, 6, 9, and 18 months only). All stats are computed across all relevant segments in the LENA daylong corpus. For e.g., MeanDurSegs\_SegLvl is the list of mean durations of each segment type in the daylong LENA data. So, for segment type SIL, indicating silence, the mean duration is computed by listing all durations of SIL segments in the LENA daylong corpus and taking the mean.

*Colnames*: UniqSegs (list of unique segment labels; FAN and MAN are treated separately and not as a single AN type. A similar table with FAN and MAN treated as a single AN type is generated later and is used for summary statistics presented in the Burstiness paper/pre-print), MeanDurSegs\_SegLvl, StdDurSegs\_SegLvl, MedianDurSegs\_SegLvl, MaxDurSegs\_SegLvl, MinDurSegs\_SegLvl, TotDurSegs\_DataLvl (total duration of each segment type in the entire corpus)

LdayRecDayLevelSegmentNumsSummaryStats\_No0IviMerge.csv: contains summary stats about the number of each segment type (without merging segments of the same type separated by 0 s IEIs) in a LENA daylong recording file. All stats are computed across the LENA daylong corpus across ages (3, 6, 9, and 18 months only). For e.g., MeanNumSegs\_RecLvl is the list of mean number of each segment type in a LENA daylong recording file. So, for segment type SIL, indicating silence, the mean number of segments in a LENA daylong recording file is computed by listing the total number of SIL segments in each daylong recoding file in the LENA daylong corpus and taking the mean.

*Colnames*: UniqSegs (list of unique segment labels; FAN and MAN are treated separately and not as a single AN type. A similar table with FAN and MAN treated as a single AN type is generated later and is used for summary statistics presented in the Burstiness paper/pre-print), MeanNumSegs\_RecLvl, StdNumSegs\_RecLvl, MedianNumSegs\_RecLvl, MaxNumSegs\_RecLvl, MinNumSegs\_RecLvl, TotNumSegs\_DataLvl (total number of each segment type in the entire corpus)

LdaySegsDurStruct\_No0IviMerged.mat: contains more granular information about segment durations and numbers in the LENA daylong data (including some recordings at ages other than 3, 6, 9, and 18 months; these recordings are not analysed for the Burstiness paper/pre-print and do not form part of the relevant data presented in the paper/pre-print) without merging segments of the same type separated by 0 s IEIs. The .mat file contains a MATLAB structure where each element in the structure corresponds to one daylong recording file.

*Structure field names*: SegsList (list of unique segment labels. Only the first element in the structure corresponding to the first daylong recording file has a non-empty SegsList because it is redundant to store the same info for every index in the structure), TotDurOfSegInRec\_Secs (vector containing the total duration of each segment in the daylong recording file, in seconds. This vector is indexed by the segment type), TotNumOfSegInRec (vector containing the total number of each segment in the daylong recording file. This vector is indexed by the segment type), TotDurOfRec\_Secs (scalar; the total duration of the daylong recording, in seconds), DurOfSegments (cell array containing vectors listing the durations of all segments of each segment type in the daylong recording file, in seconds. This cell array is indexed by the segment type), InfID (string; infant ID for the recording), InfAgeMos (scalar; infant age for the recording, in months)

LdaySegmentLevelDurationSummaryStats\_0IviMerged\_ChnAdOnly.csv, LdayRecDayLevelSegmentNumsSummaryStats\_0IviMerged\_ChnAdOnly.csv, LdaySegsDurStruct\_0IviMerged\_ChnAdOnly.mat: counterparts for similar files above with suffix \_No0IviMerge instead of \_0IviMerged\_ChnAdOnly. That is, these files have similar summary statistics and more granular data in a MATLAB structure file (saved as a .mat file) but only for infant (speech and non-speech related) vocalisations as well as adult vocalisations treated as a single AN label category (in line with analyses in the Burstiness paper/pre-print). Further, durations and vocalisations numbers (and summary statistics) are estimated *after* vocalisations in the same label category separated by 0 s IEIs have been merged (also in line with analyses in the Burstiness paper/pre-print). As such, these files are used to report summary statistics and other relevant plots for infant speech-related, infant non-speech-related, and adult vocalisations in the Burstiness paper/pre-print, while the files with the suffix \_No0IviMerge are used to report summary statistics for all other segment labels wherever they are provided in the Burstiness paper/pre-print. Note that summary statistics in the .csv files are reported only for data from ages 3, 6, 9, and 18 months (combined), while the .mat file contains more granular data from some recordings at ages other than 3, 6, 9, and months (in addition to all data from ages 3, 6, 9, and 18 months). Column names in the .csv files and structure field names in the .mat file are the same as those for the corresponding files with the \_No0IviMerge suffix.

LdaySegDurAndNumsSummaryStats\_OnlyChnAd0IviMerged.xlsx: This file contains the summary statistics for infant speech-related, infant non-speech-related, and adult vocalisations *after* the 0 s IEI merge has been performed as well as the summary statistics for all other segment types which are reported without the 0 s IEI merging. This file is generated by manually merging summary statistics from the relevant .csv files (LdaySegmentLevelDurationSummaryStats\_<0sIviMergeStatus\_and\_VocTypes>.csv, LdayRecDayLevelSegmentNumsSummaryStats\_<0sIviMergeStatus\_and\_VocTypes>.csv). See the Notes sheet in the file for details.

<ValidationDataType>SegmentLevelDurationSummaryStats\_0IviMerged\_ChnAdOnly.csv: validation data counterpart files to LdaySegmentLevelDurationSummaryStats\_0IviMerged\_ChnAdOnly.csv, with separate .csv files for LENA 5min data, human listener labelled data with all adult vocalisations, and human listener labelled data with only child-directed adult vocs.

<ValidationDataType>RecDayLevelSegmentNumsSummaryStats\_0IviMerged\_ChnAdOnly.csv: validation data counterpart files to LdayRecDayLevelSegmentNumsSummaryStats\_0IviMerged\_ChnAdOnly.csv, with separate .csv files for LENA 5min data, human listener labelled data with all adult vocalisations, and human listener labelled data with only child-directed adult vocs.

<ValidationDataType>SegsDurStruct\_0IviMerged\_ChnAdOnly.mat: validation data counterpart files to LdaySegsDurStruct\_0IviMerged\_ChnAdOnly.mat, with separate .mat files for LENA 5min data, human listener labelled data with all adult vocalisations, and human listener labelled data with only child-directed adult vocs.

*Structure field names*: SegsList (list of unique segment labels. Only the first element in the structure corresponding to the first daylong recording file has a non-empty SegsList because it is redundant to store the same info for every index in the structure), TotDurOfSegInRec\_Secs (vector containing the total duration of each segment in the daylong recording file, in seconds. This vector is indexed by the segment type), TotNumOfSegInRec (vector containing the total number of each segment in the daylong recording file. This vector is indexed by the segment type), TotFileDur\_Secs (scalar; the total duration of the validation data file, in seconds. This is estimated as the number of 5-minute sections times 300 s and also accounts for the one validation data file that has one 3-minute section due to human error. See supplementary text for the Burstiness paper/pre-print as well as the wiki for the OSF repository associated with the Burstiness paper/pre-print. Note that the total duration of audio tagged to be annotated is what is represented here vs. the total duration of vocalisations identified as infant or adult by the human annotator), DurOfSegments (cell array containing vectors listing the durations of all segments of each segment type in the daylong recording file, in seconds. This cell array is indexed by the segment type), InfID (string; infant ID for the recording), InfAgeMos (scalar; infant age for the recording, in months)

TotVoc<Dur/Nums>Summaries\_0IviMerged\_ChnAdOnly\_<DataType>.xlsx: contains total vocalisation duration (Dur) or total vocalisation count (Nums) summaries for infant and adult vocalisations for all datasets analysed in the Burstiness paper/pre-print. Various sheets within these .xlsx files contain total vocalisation duration or total vocalisation count summaries (as applicable) broken down at various levels. Summaries are only reported for data at 3, 6, 9, and 18 months. Summaries are not reported for the human listener-labelled data with only child-directed adult vocalisations included since this info is present in the relevant file for the human listener-labelled data with all adult vocalisations included (see below for descriptions of each sheet in the.xlsx file). Finally, summaries are reported *after* vocalisations of the same type separated by 0 s IEIs have been merged.

Tot<Dur/Num>: contains the total vocalisation durations (or total vocalisation counts, as applicable) for each speaker label (CHNSP, CHNNSP, AN) and annotation label (C, X, R, L, T, U, N; only for human-listener labelled data) summed across all files and all ages (3, 6, 9, and 18 months). Also contains the total vocalisation duration/counts for all infant (CHN; speech-related and non-speech related combined) and the total vocalisation duration/counts summed across all speaker labels (CHNSP, CHNNSP, and AN combined). *Colnames*: <Dur/Num>\_<speaker/annotation label>, <Dur/Num>\_CHN, <Dur/Num>\_Tot (the overall total infant speech-related, infant non-speech-related, and adult vocalisation duration across the entire dataset)

FileLvlTot<Dur/Num>OfSegs: contains file level total vocalisation durations (or counts) for each speaker label and annotation label as applicable. Also contains the file level total duration (or counts) for the combined CHN category and for all speaker labels combined (CHNSP, CHNNSP, and AN combined). The file name root for the daylong or validation file, infant age, and the number of sections in the file (5-minute sections for the validation data vs. the number of sub-recordings for the LENA daylong data) are also provided. *Colnames*: <Dur/Num>\_<speaker/annotation label>, <Dur/Num>\_CHN, <Dur/Num>\_Tot, FileRoot, InfantAgeMonth, NumSections

Tot<Dur/Num>ByAgeAndVocType: contains totals broken down by infant age and vocalisation (and annotation, as applicable) type, summed across all files. For eg., the sum reported for CHNSP at age 3 months is the sum of all CHNSP vocalisations across all files at 3 months. *Colnames*: AgeBlock (infant age in months), ANOrCHNLabels (totals at the level of AN and CHN labels), SpkrLvlLabels (totals at the level of AN, CHNSP, and CHNNSP labels), HAnnots (totals at the level annotation label level, if applicable), Totals

Tot<Dur/Num>ByVocType: contains totals broken down by vocalisation and annotation type, as applicable, summed across all files. For eg., the sum reported for CHNSP is the sum of all CHNSP vocalisations across all files. *Colnames*: ANOrCHNLabels (totals at the level of AN and CHN labels), SpkrLvlLabels (totals at the level of AN, CHNSP, and CHNNSP labels), HAnnots (totals at the level annotation label level, if applicable), Totals

Tot<Dur/Num>ByAge: contains totals broken down by infant age, summed across all files. For eg., the sum reported for 3 months is the sum of all vocalisations across all files at 3 months. *Colnames*: AgeMonth, TotsByAge

TotAudioLengthAndKeySegPropDur.csv: contains info about the total duration of audio in seconds and hours (estimated as the total length of the daylong recording for the LENA daylong data and the total length of audio tagged to be annotated for all annotated 5-minute sections for the validation data), total duration of key segments (CHNSP, CHNNSP, AN combined) in seconds and hours, and the percent of total audio duration occupied by key segments for all data types (LENA daylong data as well as all validation datasets).

*Colnames*: DataType, TotAudioDur\_s, TotAudioDur\_hr, TotKeySegDur\_s, TotKeySegDur\_hr, PercKeySegDurInTotAudio

TotNumMerges\_NVocsTo1Voc\_ByVocType\_ReqAgesOnly.csv: summarises info about the number of vocalisation merge chains (where *n* > 2 vocalisatons of the same type separated by 0 s IEIs are merged into one vocalisation) by vocalisation type and dataset type.

*Colnames*: DataType, CHNSP (the number of vocalisations merge chains for infant speech-related vocalisations for each dataset), CHNNSP (the number of vocalisations merge chains for infant non-speech-related vocalisations for each dataset), AN (the number of vocalisations merge chains for adult vocalisations for each dataset)

**ReliabilityTabs**: contains reliability results.

ReliabilityErrorRates\_FileOrSectionLvl.csv: summarises (at the file level and the 5-minute section level) reliability numbers (Cohen’s kappa, percent agreement, number of false alarms, number of misses, number of confusions), number of 1 ms frames that LENA and the human listener agree on the same label for each label type, the number of 1 ms frames that are labelled a given type by LENA for each label type, and the number of 1 ms frames that are labelled a given type by the human listener for each label type. These numbers are provided for LENA 5-min data vs. human listener-labelled data with all adult vocalisations included and only infant-directed adult vocalisations included.

*Colnames*: NumFrameMatch\_<LabelType> (number of 1 ms frames LENA and the human listener agree as a given label type), NumFramesLENA\_<LabelType> (number of 1 ms frames LENA identifies as a given label type), NumFramesHum\_<LabelType> (number of 1 ms frames the human listener identify as a given label type), NumFA (number of 1 ms frames that are false alarms), NumMiss (number of 1 ms frames that are misses), NumConf (number of 1 ms frames that are confusions when CHNSP and CHNNSP are considered as separate labels), NumConfCHN, NumSpeech\_Hum (number of 1 ms frames that are confusions when CHNSP and CHNNSP are considered as a single CHN label), CohensKappa, PercentAgreement, Fname (file name), AgeMonths, SecOrFileLvl (whether the estimated numbers are at the validation file level or the 5-minute section level), AN\_Type (all adult vocalisations included or only infant-directed adult vocalisations included)

ConfusionMatStruct.mat: contains cell arrays containing confusion matrices (see Fig. S18B in the supplementary info of the Burstiness paper/pre-print) as well as precision row vectors and recall column vectors for each 5-minute section and validation data file, and for human listener-labelled data with all adult vocalisations included and only infant-directed adult vocalisations included. The combination of section/file level and all adult vocalisations/infant-directed adult vocalisations is used to index the structure such that there are 4 elements in the structure, one for each combination of section/file level and all adult vocalisations/infant-directed adult vocalisations. For eg., for the section level + all adult vocalisations combo, the structure field corresponding to the cell array containing confusion matrices has 420 elements corresponding to each of the 420 annotated 5-minute sections in the validation data. The confusion matrices in this case are computed based on the human listener-labelled dataset with all adult vocalisations included.

*Structure field names*: LabelSetIntersectNums\_Cell (cell arrays containing confusion matrices), NumDenom\_Prec\_RowVec\_Cell (cell array containing precision row vectors), NumDenom\_Recall\_ColVec\_Cell (cell array containing recall column vectors), HumAdultVocType (all adult vocalisations included or only infant-directed adult vocalisations), SecOrFileLvl (5-minute section or validation data file level)

ReliabilityAggregateTab.xlsx: contains sheets with mean and total reliability numbers for LENA 5-min validation data when compared against human listener-labelled data with all adult vocalisations included (H\_AllAd\_FileLvl) and with only infant-directed adult vocalisations included (H\_TAd\_FileLvl). Totals are computed for the entire validation dataset and mean values are calculated based on file-level estimates of reliability metrics.

*Colnames*: Vars (reliability metric names), MeanVals (mean reliability metric values, averaged over reliability metrics computed at the validation data file level. For eg., the mean false alarm rate is computed by averaging over the false alarm rates for each validation data file), TotVals (total reliability metrics values. For eg., the total false alarm rate is computed by dividing the total number of false alarm frames in the dataset by the total number of frames the human listener identifies as speech)

**UttDurTabs**: contains .csv files with utterance durations for different data types.

*File name format*: Durations\_<DataType>.csv

*Colnames*: Duration, Speaker, InfAgeMnth, FnameUnmerged (name of the sub-recording the duration is from)