Segment Extraction Supplemental Material

The LENA ".its" file is an xml file that provides information about the LENA recorder version, software version, information about the automatically-produced sound type labels, and much other information. Here is an example of a portion of one of the .its files used in this study, corresponding to Figure 2 in the main article:

<Segment spkr="CHN" average_dB="-26.01" peak_dB="-18.84" conversationInfo="|RC|10|9|3|AICF|NT|FH|" childUttCnt="1" childUttLen="P1.55S" startUtt1="PT287.69S" endUtt1="PT289.24S" childCryVfxLen="P0.00S" startTime="PT287.69S" endTime="PT289.24S" />
<Segment spkr="FAN" average_dB="-43.54" peak_dB="-35.31" conversationInfo="|RC|10|9|3|AICF|TIFI|FI|" femaleAdultWordCnt="5.28" femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="PT289.24S" endTime="PT290.28S" />
<Segment spkr="CHN" average_dB="-45.82" peak_dB="-36.08" conversationInfo="|RC|10|10|4|AICF|TIFR|FI|" childUttCnt="1" childUttLen="P0.60S" startUtt1="PT290.28S" endUtt1="PT290.88S" childCryVfxLen="P0.00S" startTime="PT290.28S" endTime="PT290.88S" />
<Segment spkr="FAN" average_dB="-40.63" peak_dB="-31.29" conversationInfo="|RC|10|10|4|AICF|TIFE|FI|" femaleAdultWordCnt="3.49" femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="PT290.88S" endTime="PT292.08S" />

<Segment spkr="FAN" average_dB="-39.64" peak_dB="-31.73" conversationInfo="|RC|10|10|4|AICF|NT|FH|" femaleAdultWordCnt="8.86" femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="PT293.65S" endTime="PT295.66S" />

<Segment spkr="FAN" average_dB="-39.58" peak_dB="-31.36" conversationInfo="|RC|10|10|4|AICF|TIFI|FH|" femaleAdultWordCnt="7.87" femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S" startTime="PT298.31S" endTime="PT299.98S" />

<Segment spkr="CHN" average_dB="-16.19" peak_dB="-6.59" conversationInfo="|RC|10|11|5|AICF|TIFR|FI|" childUttCnt="1" childUttLen="P2.61S"

startUtt1="PT300.85S" endUtt1="PT303.46S" childCryVfxLen="P0.00S"
startTime="PT300.85S" endTime="PT303.69S" />
<Segment spkr="TVF" average_dB="-57.17" peak_dB="-45.99" startTime="PT303.69S"
endTime="PT304.78S" />
<Segment spkr="SIL" average_dB="-57.76" peak_dB="-49.54" startTime="PT304.78S"
endTime="PT307.39S" />
<Segment spkr="CHN" average_dB="-25.74" peak_dB="-17.37"
conversationInfo="|RC|10|11|5|AICF|NT|FH|" childUttCnt="2" childUttLen="P2.21S"
startUtt1="PT307.39S" endUtt1="PT307.74S" startUtt2="PT308.41S" endUtt2="PT310.27S"
childCryVfxLen="P0.00S" startTime="PT307.39S" endTime="PT310.27S" />
<Segment spkr="NOF" average_dB="-50.92" peak_dB="-44.79"
startTime="PT310.27S" endTime="PT311.10S" />
<Segment spkr="FAN" average_dB="-43.20" peak_dB="-31.10"
conversationInfo="|RC|10|11|5|AICF|TIFI|FI|" femaleAdultWordCnt="5.68"
femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S"
startTime="PT311.10S" endTime="PT312.76S" />
<Segment spkr="CHN" average_dB="-28.98" peak_dB="-18.43"
conversationInfo="|RC|10|12|6|AICF|TIFR|FI|" childUttCnt="1" childUttLen="P0.99S"
startUtt1="PT312.76S" endUtt1="PT313.75S" childCryVfxLen="P0.00S"
startTime="PT312.76S" endTime="PT313.75S" />
<Segment spkr="OLN" average_dB="-26.88" peak_dB="-23.08"
startTime="PT313.75S" endTime="PT314.55S" />
<Segment spkr="FAN" average_dB="-36.17" peak_dB="-29.04"
conversationInfo="|RC|10|12|6|AICF|TIFE|FI|" femaleAdultWordCnt="6.74"
femaleAdultNonSpeechLen="P0.00S" femaleAdultUttCnt="0" femaleAdultUttLen="P0.00S"
startTime="PT314.55S" endTime="PT315.65S" />
<Segment spkr="MAN" average_dB="-46.61" peak_dB="-39.19"
conversationInfo="|EC|10|12|6|AICF|NT|FI|" maleAdultWordCnt="4.82"
maleAdultNonSpeechLen="P0.00S" maleAdultUttCnt="0" maleAdultUttLen="P0.00S"
startTime="PT315.65S" endTime="PT316.95S" />

</Conversation>

<Pause num="11" average_dB="-45.86" peak_dB="-28.31" childCryVfxLen="P0.00S"
femaleAdultNonSpeechLen="P0.00S" maleAdultNonSpeechLen="P0.00S" TVF="P0.00S"
FAN="P0.00S" OLN="P0.00S" SIL="P2.97S" NOF="P16.24S" CXF="P0.00S" OLF="P0.00S"
CHF="P0.00S" MAF="P1.00S" TVN="P0.00S" NON="P0.00S" CXN="P0.00S" CHN="P0.00S"
MAN="P0.00S" FAF="P0.00S" startTime="PT316.95S" endTime="PT337.16S" >

<Segment spkr="SIL" average_dB="-57.86" peak_dB="-52.58" startTime="PT316.95S"
endTime="PT318.05S" />

We parsed this text line-by-line in Perl (code and output are provided in the *DataAndScripts Supplementary*), using regular expressions to search for sound types (*Segment spkr="CHN"* indicates a segment where the sound source is the child wearing the recorder, *Segment spkr="FAN"* indicates a segment where the sound source is an adult female, and *Segment spkr="MAN"* indicates a segment where the sound source is

an adult male). The *childUttLen* variable gives the amount of time within a child segment that is thought to be speech-related vocalization. If *childUttLen* was greater than zero we took the child vocalization segment to be speech-related; if *childUttLen* was equal to zero we took the child vocalization segment to be not speech-related. The *startTime* and *endTime* variables give the start and end times, respectively, of each speaker segment in seconds into the recording.

We treated sound types ending in *F* as if they were silence, since an *F* indicates that the labeling algorithm found the segment to have a high level of similarity to its Gaussian mixture model for silence. The bottom portion of the .its file provides information about the bar plot data in the LENA software GUI; we ignored this portion.

Segments labeled as having overlap between a human and any other sound source were excluded. It is likely that segments labeled as "Overlap" do often contain adult or child vocalizations. However, because of limitations of the automated labeling system, it was not possible to determine what combination of speakers and other sounds were represented in a given Overlap segment. For this reason, we opted to focus on cases where the speaker identities were known. It could be argued that for children's learning there might be advantages for vocalizations produced with a background of silence compared to vocalizations produced in overlap with another sound source, but that needs to be empirically determined.

Reliability of the automated labeling compared to human listener labeling has been computed for recordings of typically developing children, yielding the following agreement with human transcribers: 82% for adult speaker, 76% for child wearing the

recorder, 75% for child speech-like utterances, and 84% for child non-speech-like vocalizations (Xu, Yapanel, & Gray, 2009).

Human listener labeling was not done for the recordings of children with ASD. The training of the Gaussian mixture models that labeled all the utterances in both TD and ASD samples was conducted with over 200 hours of human listener labeling on TD infant and child vocalizations as their standard. Thus, we reason, if the vocalizations of the ASD children were treated differently by the automated labeling, acoustic differences in the vocalizations produced by the children with ASD likely formed the bases for such differences.

References

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA™ language environment analysis system in young children's natural home environment* (LTR-05-2). Boulder, CO: LENA Foundation. Retrieved from http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2