LENA: Automated Analysis Algorithms and Segmentation Detail: How to interpret and not overinterpret the LENA labelings

D. Kimbrough Oller
The University of Memphis, Memphis, TN, USA and
The Konrad Lorenz Institute
for Evolution and Cognition Research (KLI), Altenberg, Austria

Supported by NIDCD, NICHD, the KLI,
and by the Plough Foundation

Overview of goal

- The software of LENA yields a very useful labeling
- The proof of its value is in outcomes (prediction of age, group classification, correlation with other language measures)
- So at a global level the system has proven itself and more importantly
- It has proven that automated analysis of massive samples is here to stay

Interpretive subtlety as a key to the long-term value of the approach

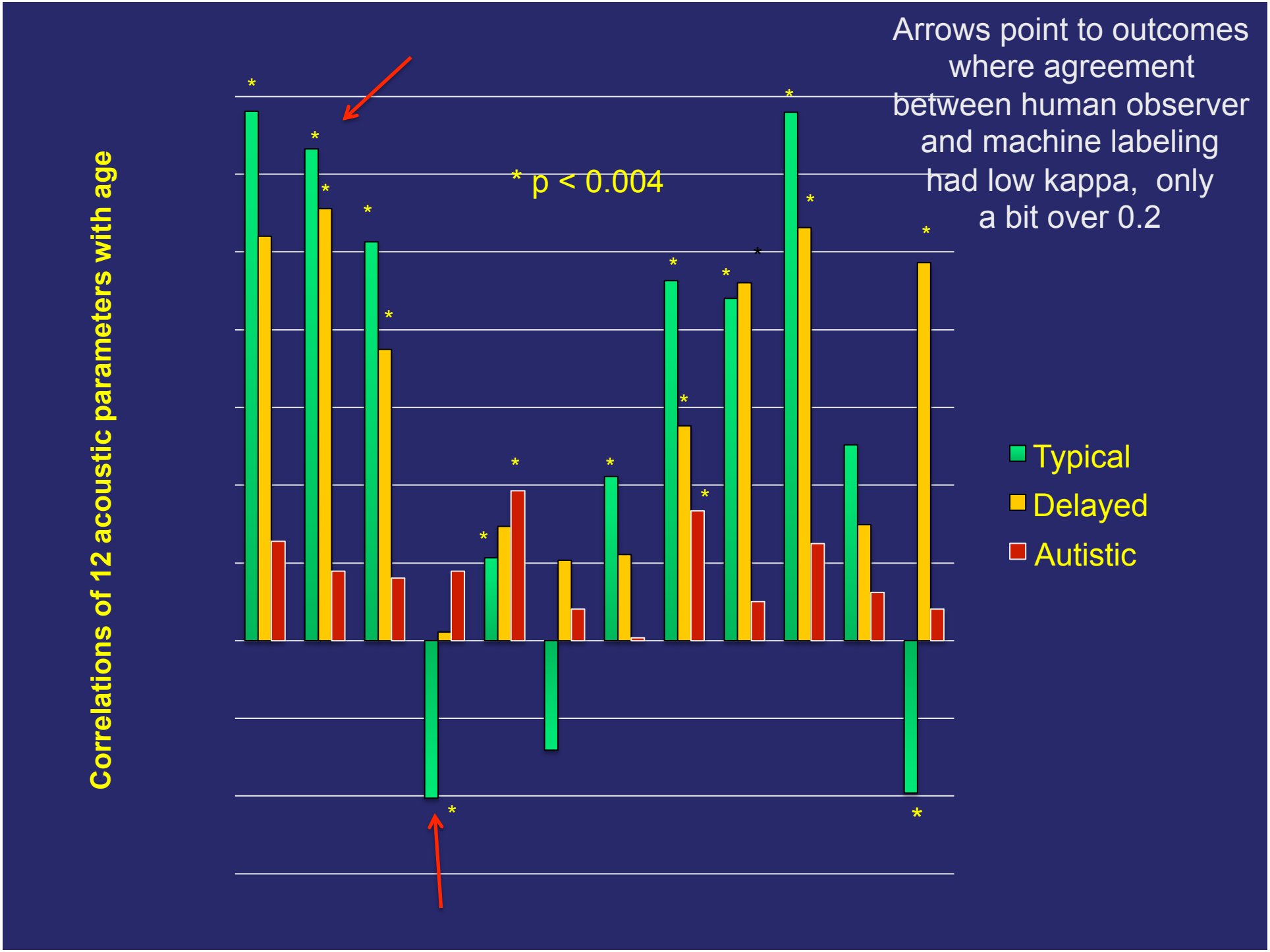
- We're going to focus on the labeling functions and how to interpret their outcomes appropriately
- The methods are designed to yield a maximally accurate outcome at a global level – the level of the recording
- The labeling at the local level is subordinated to this global accuracy goal
- Much of what one sees in a real labeled file is not correct

The key conclusions

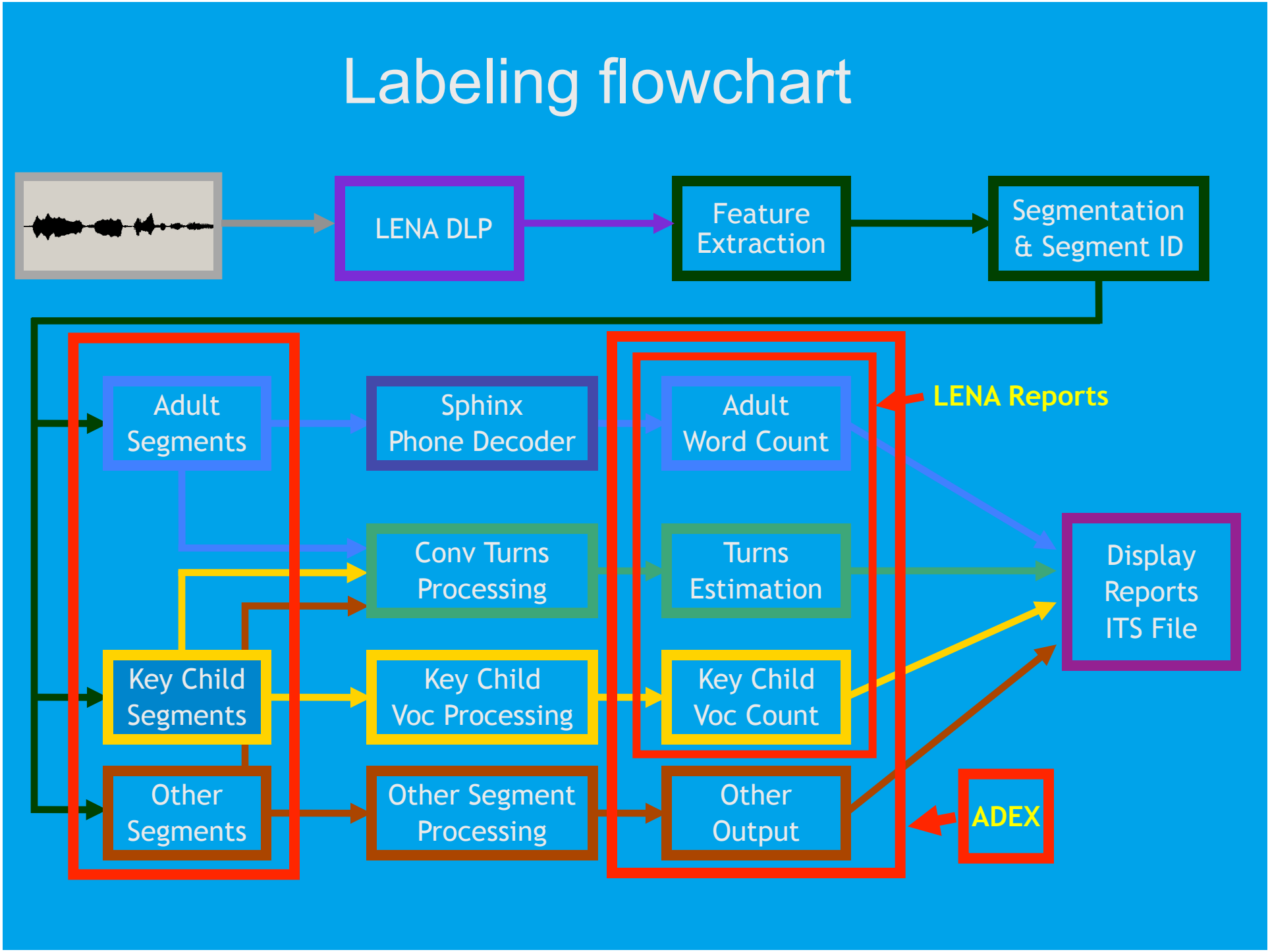
- The many mistakes that the software makes at the local level require us to be intelligent about how we use the information
- We need to think about ways the software might lead us astray
- But at the same time we need to capitalize on the opportunities of the new method and not be swayed by irrelevant traditional thinking that insists on some arbitrary metric of reliability

Maintain optimism

- E.g., low kappa is not necessarily a reason to discard data on any particular automated coding
- One needs always to look at the outcome comparisons and reason again about the significance of a low reliability factor
- I envision a back and forth between modeling and various outcomes
- In the following graph based on data from the PNAS paper, the canonical syllable (CS) and squeal (SQ) parameters (see red arrows) had very low (but positive and highly statistically significant) kappas of agreeement, yet they were strong predictors of age and of group differentiation


Correlations of 12 acoustic parameters with age
* p < 0.004
Arrows point to outcomes where agreement between human observer and machine labeling had low kappa, only a bit over 0.2
Typical
Delayed
Autistic

Labeling flowchart



Segmentation or “labelling” topics

- There are eight basic categories of “segment” or label
- The Near/Far distinction (based on a likelihood ratio test where SIL likelihood is the denominator) yields seven additional categories; thus 15 total categories
- Within key child, additional distinctions
  - Childvoc (or SCU, the term used in the PNAS paper)
  - Cry
  - Veg and fixed signals other than cry (including laugh) : VegFix

Gaussian mixture models (GMMs) at the core of the labeling

- Imagine eight acoustic representations (GMMs), all random noise at the beginning of training, each with the task of learning to resemble the acoustic characteristics of one of the 8 basic categories
- Imagine that a GMM is presented with segments that have been labeled by human transcribers as the category it is supposed to model, and that on each presentation, the GMM makes an adjustment in its acoustic characteristics to bring it a little closer to the characteristics of the presented segment
- All the 8 GMMs get this kind of training
- After very large numbers of presentations of labeled segments each GMM tends to stabilize as a model of the kind of segment it is supposed to model

More on the Gaussian mixture models (GMMs)

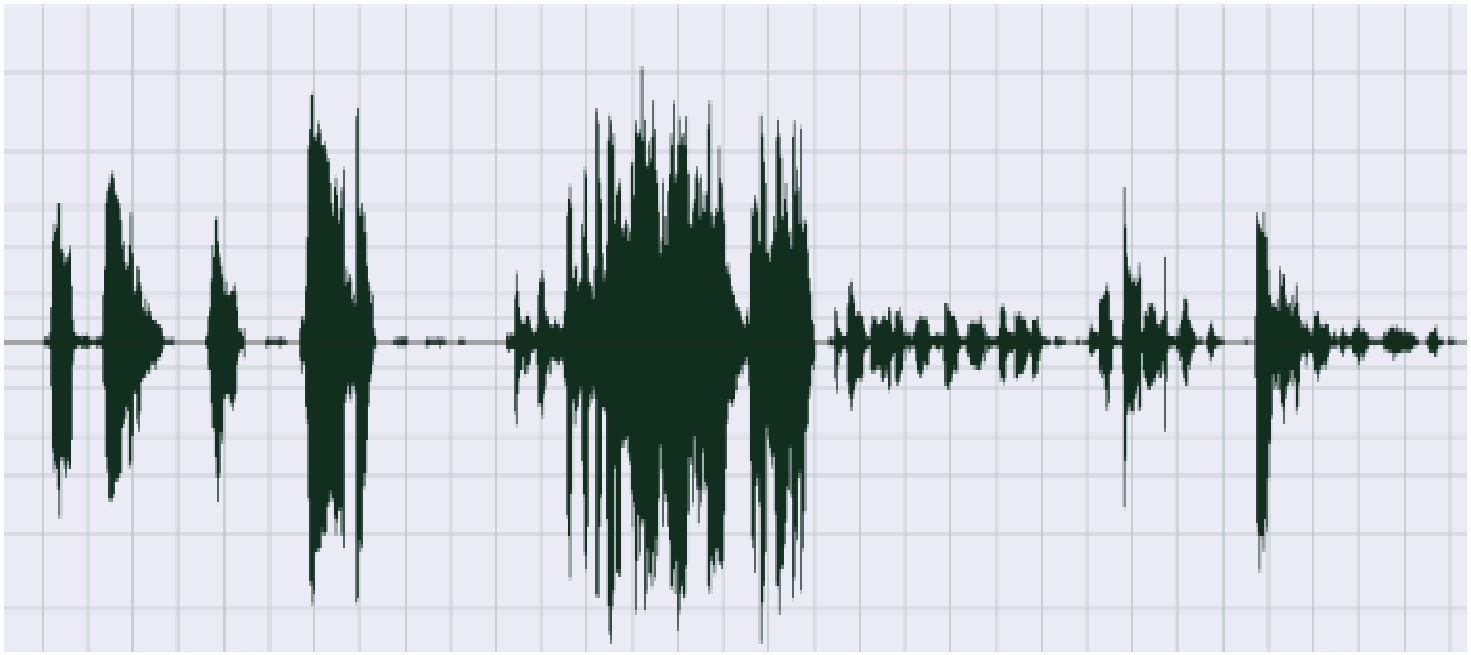
- After training, all the GMMs are non-random, each a composite model of its category (one has acoustic properties of Female Adult utterances, one of Male Adult and so on), based on many different exemplars that had been presented in training
- To test for reliability of the GMMs, they are presented with new human-labeled segments, that had not been involved in the model training, and the machine labeling is compared quantitatively with the human labeling

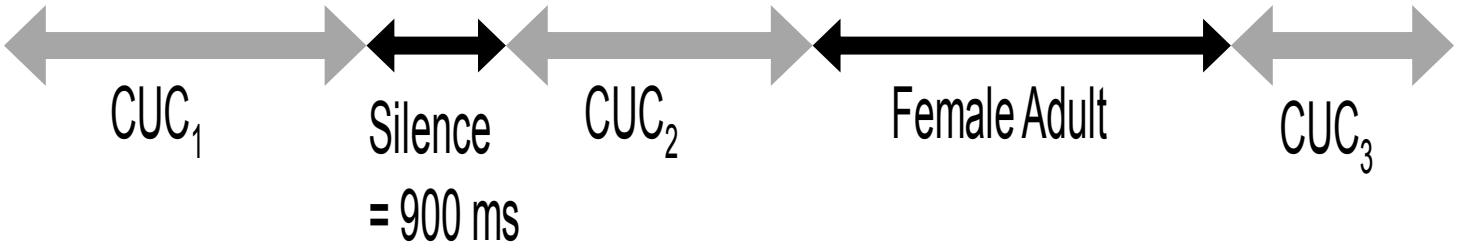
Labeling constraints

- Min duration constraints on labeled events
  - 1000 ms for MAN/FAN/TVN/OLN
  - 800 ms for SIL, NON, CXN
  - 600 ms for CHN
- The special category of Overlap (OLN/OLF); must include a voice, but remember, it is based on its own GMM where training exemplars included one or more voices plus possible other sounds
- The start and end times of labeled events are often not where a listener would place them (30-40 ms errors are common)
- Vocal activity blocks (VABs) and the related idea of Conversations vs Pauses
  - the 5 sec rule is used for boundaries between VABs

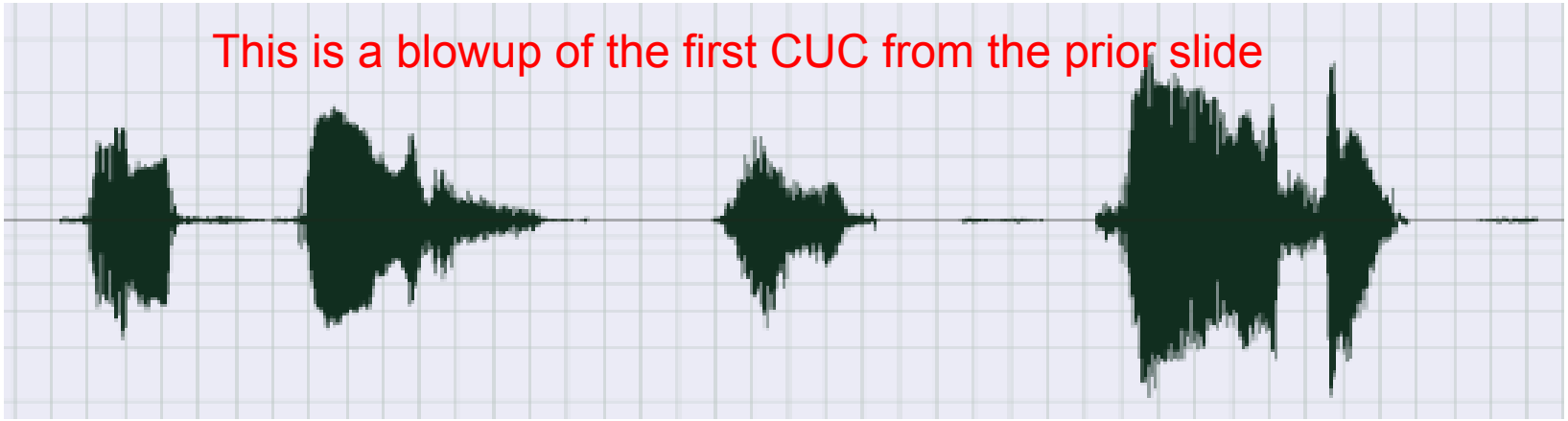
Other durational constraints

- Child vocalizations (Childvoc) within CHN/CHF begin when the acoustic energy level first rises to 90% above baseline for at least 50 ms and end when it falls to less than 10% above baseline for at least 300 ms
- Thus 50 ms is the absolute min for a Childvoc, and 300 ms is the max break within a Childvoc
- The easier way to think about this may be that Childvocs are never too short, and never broken up by long silences (never broken up by a silence as long as a typical syllable, i.e. 300 ms), but can consist of long utterances with many syllables
- When a silence longer than 300 ms occurs within a CHN/CHF, a new Childvoc begins

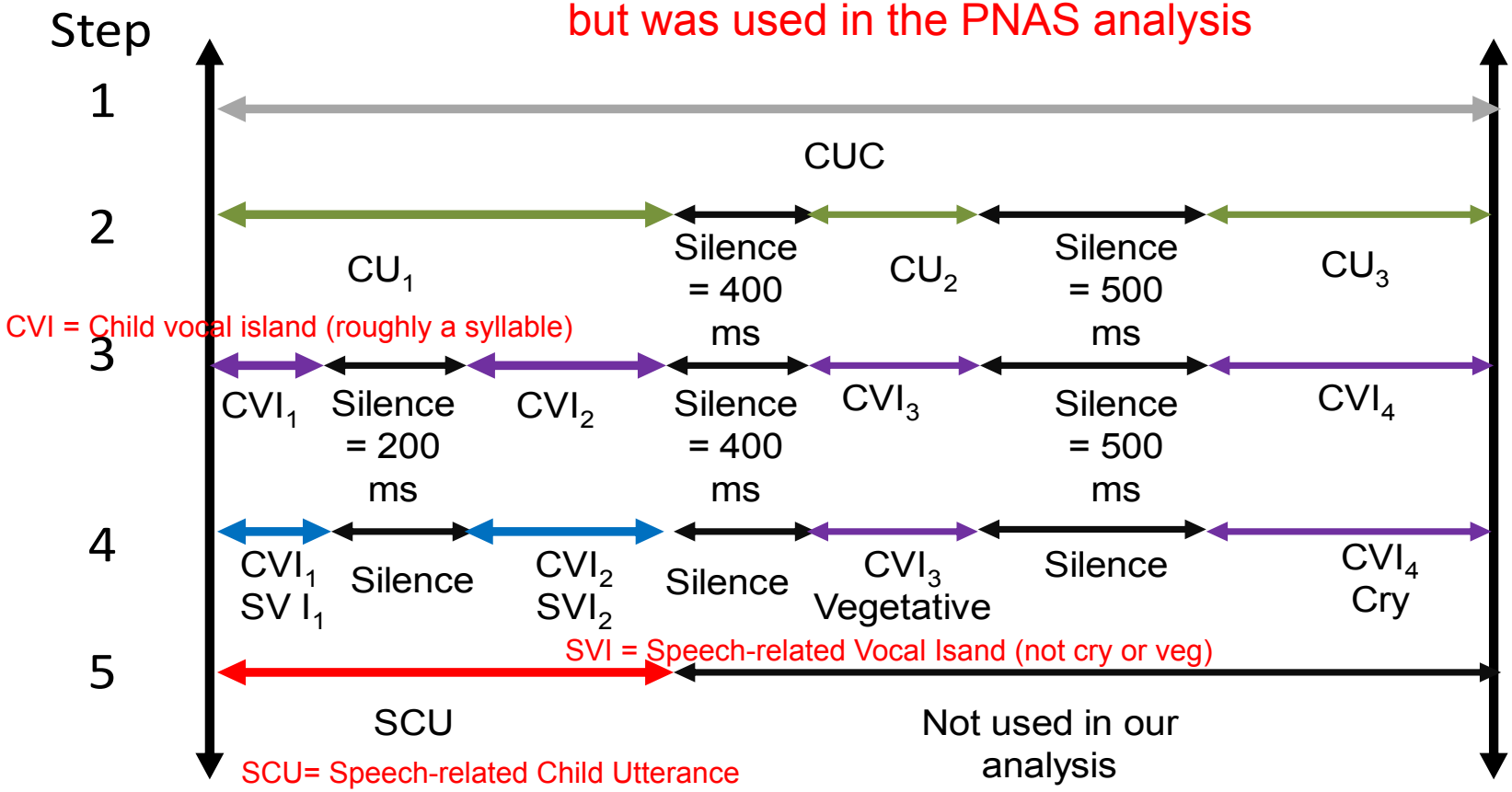




CUC= child utterance cluster or CHN/CHF



Vocal Island analysis is not a part of the standard LENA algorithms, but was used in the PNAS analysis



How was reliability of segmentations assessed?

- 70 hours of transcribed data in 6 ten-min chunks from each of 70 children balanced for gender and age were used for testing
- This was done with the segmentations from the automated system in front of the transcribers (in the open source software “Transcriber”)
- Transcribers moved boundaries and relabeled with many more categories than the 15 (>70)
- The lead transcriber reviewed every segment in the entire 70 hours before submission to reliability tests
- Transcribers were encouraged to be critical of the machine labeling
- Often transcriptions showed events violating the min duration constraints

How reliable are the segmentations?

- The comparison between machine and transcribers was done at the frame level (10 ms)
- Collar guard at various settings (nominal 30 ms, the value used for the PNAS paper) to allow small errors without penalty at the start and end times of segments
- These yield over 0.7 agreement in most cells of the reliability matrices (many have been computed)

a. When rows sum to one , the human listener is the gold standard

Human listener classification	Machine classification	
	Key child	Other
Key child	0.73	0.27
Other	0.05	0.95

b. When columns sum to one, the machine is the gold standard

Human listener classification	Machine classification	
	Key child	Other
Key child	0.64	0.03
Other	0.36	0.97

These data give a picture of the accuracy of the algorithms within the CHNs and CHFs, that is the accuracy of differentiation of Speech related material from cries and vegetative sounds

a.

Human listener classification	Machine classification	
	SVI	Cry/Vegetative
SVI	0.75	0.25
Cry/Vegetative	0.16	0.84

b.

Human listener classification	Machine classification	
	SVI	Cry/Vegetative
SVI	0.86	0.28
Cry/Vegetative	0.14	0.72

What is a Vocal Activity Block (or conversation)?

- Lots of room for containing a variety of event types
- Must contain at least one of the following 4 segment types, or any combination of them: MAN, FAN, CHN, or CXN
- But a VAB can be broken up by (i.e., a new VAB starts at) any combination of more than 5 sec of the 11 segment types that cannot be part of a “conversation”, namely MAF or FAF or CHF or CXF or OLF or OLN or NOF or NON or TVF or TVN or SIL
- And of course a VAB can include within it, any combination of ***less*** than 5 sec of the segments that cannot be part of a “conversation”

What is a conversational turn?

- Lots of room for containing a variety of event types, but CXN is not included
- MAN or FAN + CHN in either order, within vocal activity block (VAB)
- Must not include any combination of more than 5 sec of MAF or FAF or CHF or CXF or OLF or OLN or NOF or NON or TVF or TVN or SIL
- AND, if a CXN intervenes between a MAN or FAN + CHN in either order, no conversational turn is counted
- A FINAL CONSTRAINT: A conversational turn is invalidated by any FAN or MAN that was given a 0 word count by the word count module (AWC)

In summary, what is a pause between VABs?

- Lots of room for containing a variety of event types
- Any Far event, OLN or TVN, NON, or SIL
- Must consist of these things in any combination of at least 5 sec

Major things to look out for

- Reliability of labeling is pretty good at the event (segment) level
- At the level of conversational turn or any sequence of events, you reduce the reliability by amounts unknown, perhaps as much as the product of the reliabilities for the two segments (e.g., 0.7 * 0.7= 0.49)
- And consider complications of interpretation if there is overlap or far segments embedded in the turn (which they are allowed to be), or FAN/MAN with 0 word count

More technical topics

- Gaussian mixture models were trained on 230 hours of human coded data
- Labeling is based on a maximum likelihood model (for every segment in a recoding, a likelihood is determined for each of the GMMs, and the highest likelihood is chosen as the label)
- Time frame of operation of the GMMs is 10 ms, but the label decisions are made based on min length of events (i.e., twice the min length constraint, or 1200-2000 ms, is the search space)
- Iteration of procedures occurs in several instances
- TV detection is refined in subsequent passes of processing