Fidelity of LENA vocal activity labels

Mark VanDam[1] & Noah H. Silbert[2]
[1]Washington State University, [2]University of Maryland

mark.vandam@wsu.edu www.vanDamMark.com

LENA International Conference 2013
Denver, CO | 28 April, 2013

R01DC006681
P30DC04662
T32DC00013

Main research questions

1. Are LENA segment labels more accurate for children or adults?
2. Are LENA segment labels more accurate for mothers or fathers?
3. What are the implications of differences among label success rates?

Background

Automatic speech recognition (ASR)

The goal of ASR technology is (a) to segment and (b) to assign a label to each segment. Performance of segmenting and labeling can be independently evaluated.

Performance of LENA ASR

ASR agreement for segments humans labeled as
‘adult’ = 82%[1-4]
‘child’ = 76%[1-4] & 73%[5]

Human agreement for segments ASR labeled as
‘adult’ = 68%[3-4]
‘child’ = 70%[3-4] & 64%[5]

References

1. Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA Language Environment Analysis system in young children's natural home environment*. (Technical Report LTR-05-2). Retrieved from: http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2.
2. Christakis, D. A., Gilkerson, J., Richards, J. A., Zimmerman, F. J., Garrison, M. M., Xu, D., Gray, S., & Yapanel, U. (2009). Audible television and decreased adult words, infant vocalizations, and conversational turns. *Archives of Pediatric and Adolescent Medicine* 163(6): 554-558. doi:10.1001/archpediatrics.2009.61
3. Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, *40*, 555–569. doi:10.1007/s10803-009-0902-5
4. Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, 124(1): 342-349. doi: 10.1542/peds.2008-2267
5. Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences* 107(30): 13354-13359. doi:10.1073/pnas.1003882107

Method

Materials

- ► 26 families gave whole-day recordings
- ► *KEY CHILD*: TD, M=2.07 yrs (SD=.69 yrs)
- ► 30 segments from each **CHN**, **FAN**, **MAN** were excised. Segments were equally distributed throughout the day (90 segments per family)
- ► 2340 total segments

Judges

- ► 21 judges with experience in child language were recruited. Judges were students, SLPs, professors, or researchers. None reported hearing or speech perception issues, and all understood the task.

Procedure, task

- ► All judges listened to all segments; segments were individually randomized for each judge; each judge contributed about 2 hours.
- ► 4AFC task to ID “child,” “mother,” “father,” or “other.”
- ► Auditory-only playback at user-controlled volume; unlimited playback per trial
- ► minimal instructions: “These are real audio segments recorded in natural family situations. Listen to each segment, and enter the best label: 1=Child, 2=Mother, 3=Father, 4=not(Child, Mother, Father).”
- ► no feedback to judges

Data analysis

- ► 21 judges completed 2340 trials each
- ► 99.91% valid response (49097÷49140)

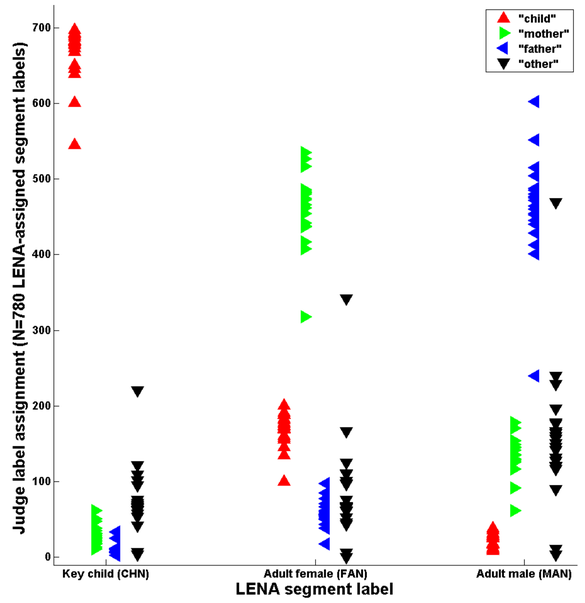

Figure 1. Judges **child**, **mother**, **father**, and **other** responses to audio segments labeled by LENA software (on the abscissa) as Key child (**CHN**), Adult female (**FAN**), and Adult male (**MAN**). Markers represent individual judges. 780 segments of each LENA label were evaluated by each judge.

Results

1. **Percent agreement between machine-coded segments and human judges and mean Kappa values:**

	%	κ	
child	**85.9**	**.708**	***all groups
mother	**59.4**	**.503**	are mutually
father	**60.9**	**.599**	distinct (p<.001)
pooled	**68.7**	**.559**	

2. **Machine-labeled *mother* tokens that were given *child* labels by human judges were the most frequent errors.**
3. **Tokens receiving each machine label were frequently labeled *other* by human judges.**

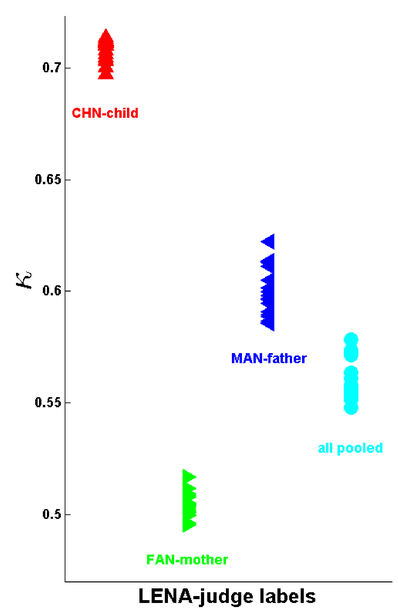

Figure 2. Kappa statistics (κ) for agreements between LENA labels and judge identifications for CHN-**child**, FAN-**mother**, MAN-**father**, and **all pooled**. Two-tailed, paired sign tests reveal mutually exclusive performance for all categories (all *p*s < .001).

Conclusions

1. **Results here are convergent with previous findings in the literature, but there are important performance differences between groups.**
2. **Machine performance is best for children, and better for adult males than adult females. Since mothers are obviously important, this is an area to improve algorithm performance.**
3. **Human-labeled *other* segments are fairly frequently given live human vocal labels by the ASR algorithm. This could profitably be improved.**