Automatic speech recognition of naturalistic recordings in families with children who are hard of hearing

Mark VanDam[1] & Noah Silbert[2]

[1]Department of Speech & Hearing Sciences
Washington State University
[2]Department of Communication Sciences & Disorders
University of Cincinnati

Talk 4pSC presented at the *167th Meeting of the Acoustical Society of America*, Providence, RI, May 8, 2014

Acknowledgments

Mary Pat Moeller
Nicholas A. Smith
Sophie E. Ambrose
Lisa Burton
J. Bruce Tomblin
Rick Arenas
Jake Oleson
D. Kim Oller
Sandie Bass-Ringdahl
Jill Gilkerson
Dongxin Xu
Jeff Richards
Shana Bailey
Elisa Romaneschi

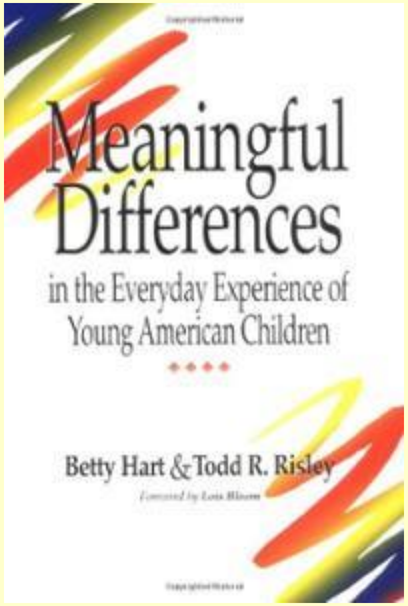
WASHINGTON STATE UNIVERSITY SPOKANE

UNIVERSITY OF Cincinnati

NIDCD

NIDCD: R01 DC009569, DC009560-01S1
PIs: MP Moeller and JB Tomblin

Hart & Risley (1995) collected child speech data in natural, home environments of 42 families.



Such data is very expensive to collect and difficult to analyze and interpret.
It took H&R 3 yrs to collect and 6 yrs to interpret.

But now, useful child language data is collected using automatic speech processing (ASP) technology.

Researchers are using the
Language ENvironment Analysis, LENA

Zimmerman, etal (2009) *Pediatrics*
Christakis, etal (2009) *Arch Pediatrics & Adol Med*
Oller, etal (2010) *PNAS*
Warren, etal (2010) *Journal Autism & Devel Disord*
Caskey, etal (2011) *Pediatrics*
Dykstra, etal (2012) *Journal of Autism*
VanDam, etal (2012) *Journal Deaf Studies & Deaf Educ*
Aragon & Yoshinaga-Itano (2012) *Sem Speech & Lang*
Suskind, etal (2013) *Comm Disord Quarterly*
Weisleder & Fernald (2013) *Psych Science*
VanDam & Silbert (2013) *POMA*
Ambrose, etal (2014) *Ear & Hearing*
....

Primary research goals of this work are to (1) address goodness of *LENA* ASR technology and (2) examine performance with hard-of-hearing children.

Data collection

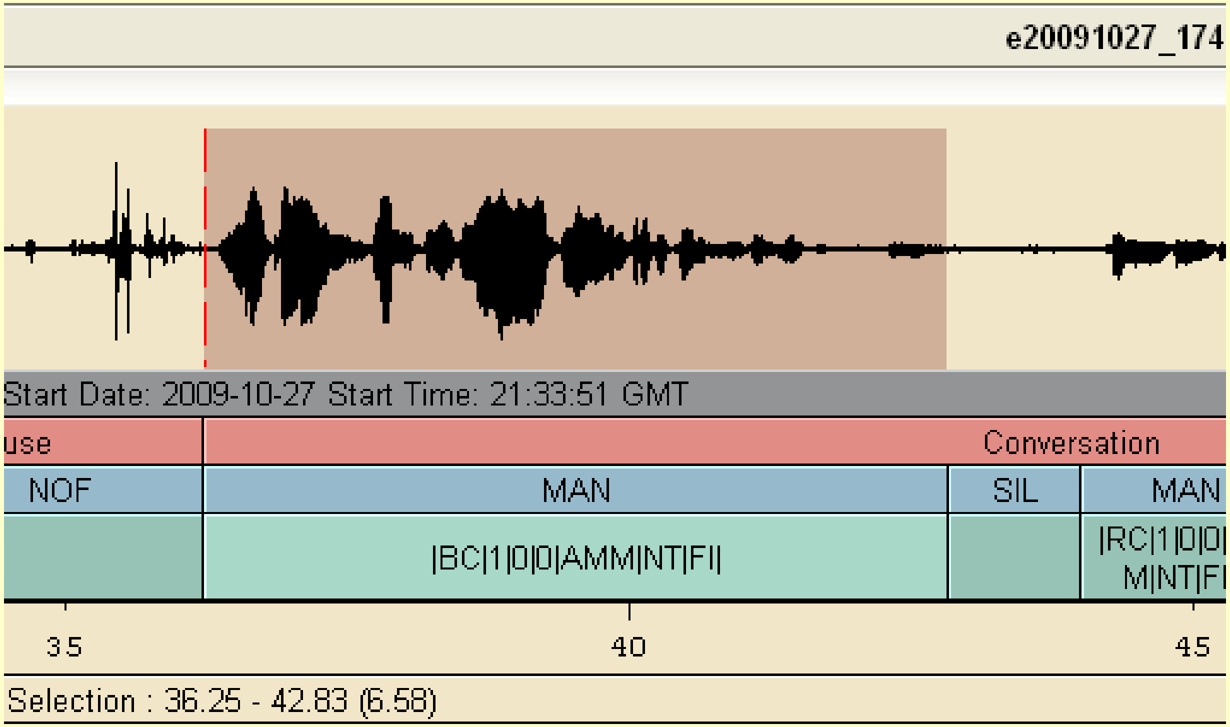
Labels on the acoustic signal:

KEY-CHILD	*OTHER-CHILD*
ADULT-MALE	*ADULT-FEMALE*

live human vocals

SILENCE	*NOISE*
ELECTRONIC (TV, RADIO)	*UNCERTAIN / FUZZY*
OVERLAPPING VOCALS	

other acoustic events



Automatic data collection results in very large database (VLDB) requiring fully automated data analyses.

Reliability of LENA labels, previous findings

ASR agreement for segments humans labeled as

'adult' = 82%

'child' = 76% & 73%

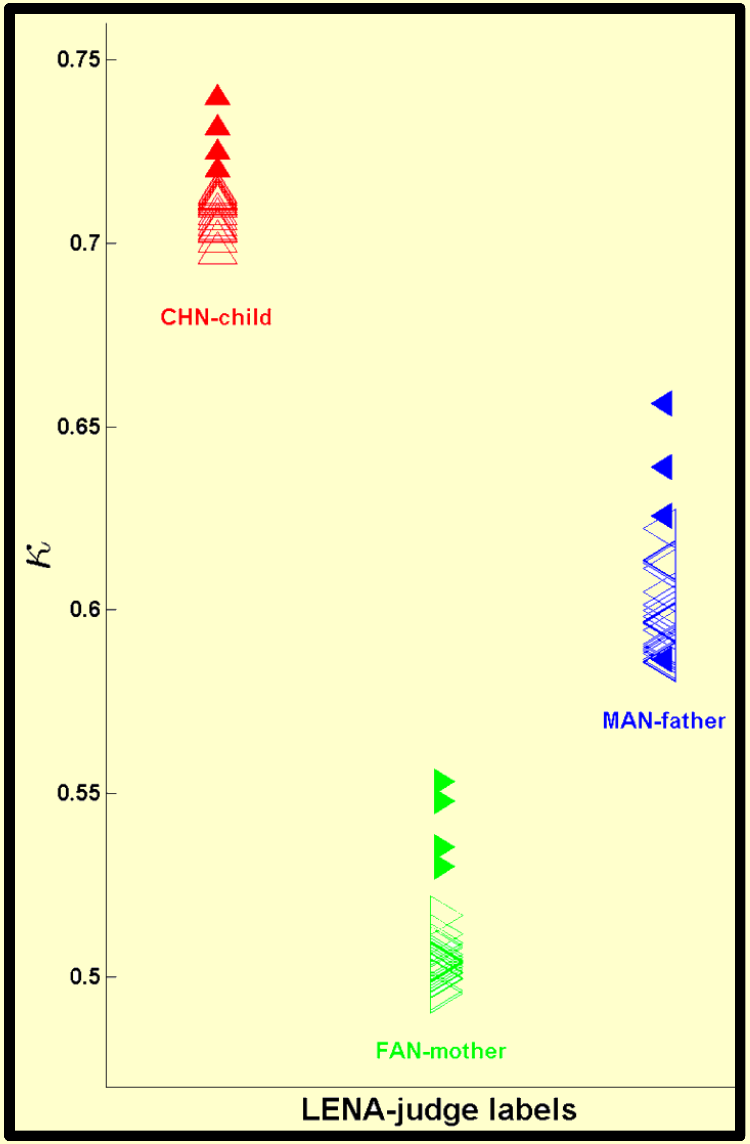
Human agreement for segments ASR labeled as

'adult' = 68%

'child' = 70% & 64%

Xu *etal* 2009; Christakis *etal* 2009; Warren *etal* 2010; Zimmerman *etal* 2009; Oller *etal* 2010

Reliability of LENA labels, previous findings



ASR—human agreement

	%	κ
CHN-child	85.9	.709
FAN-mother	59.6	.505
MAN-father	60.8	.598

Acoustic Factors:

1. duration
2. *f*0 – mean
3. *f*0 – min
4. *f*0 – max
5. *f*0 – rise
6. *f*0 – fall
7. amp, RMS
8. amp, rise
9. amp, fall
10. amp, modulate

VanDam & Silbert, 2013, *JASA/POMA*

Previous work points to duration and static F0 as primary factors in the classification.

The present work asks if classification performance changes with (speech) input from a disordered population, namely children who are hard of hearing.

The present work, method and design

-2340 tokens from 26 families with a HH child

-13 additional judges not in the first expts.

about 2hrs of listening per judge

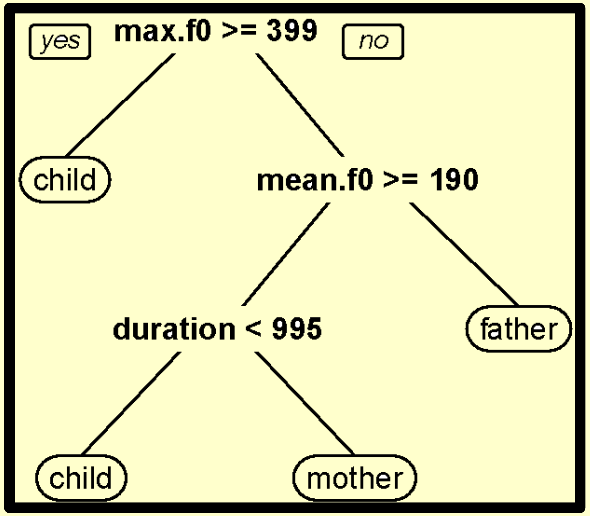
-4AFC: *mom, dad, child, other*

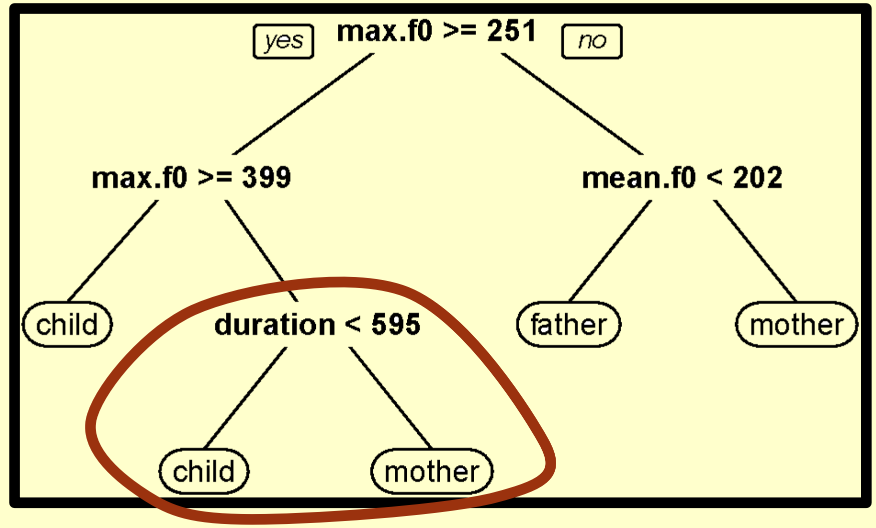
Reliability of LENA labels, current findings

How do acoustic factors drive human and machine classification decisions for families with TD kid?

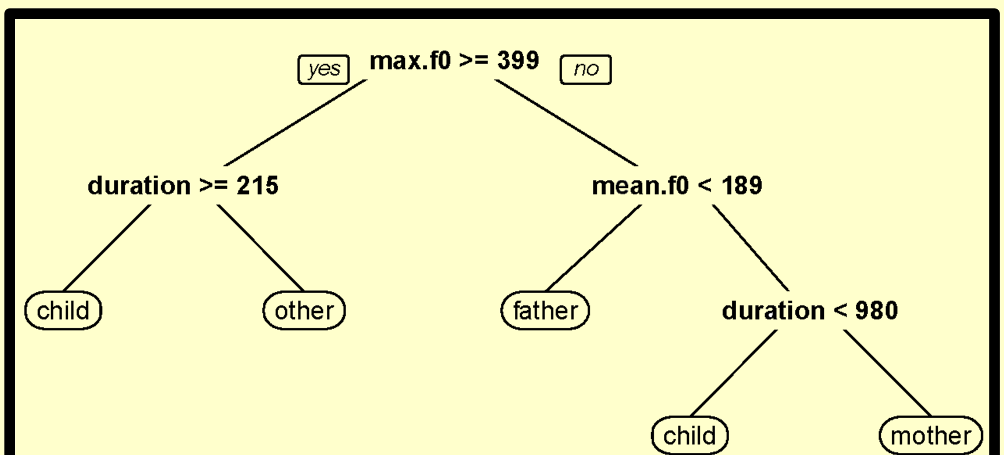
HUMANS | **MACHINE**

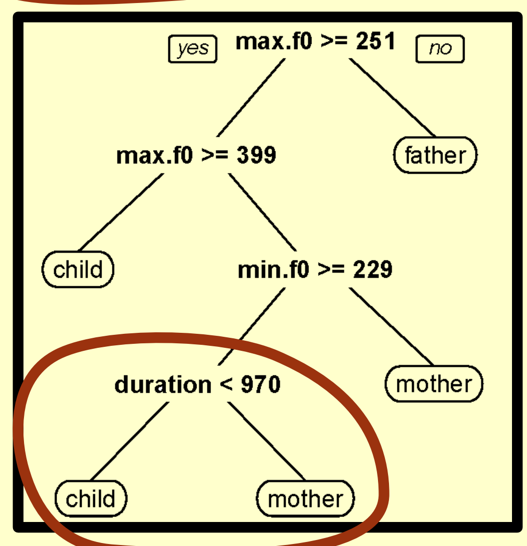
TD





HH



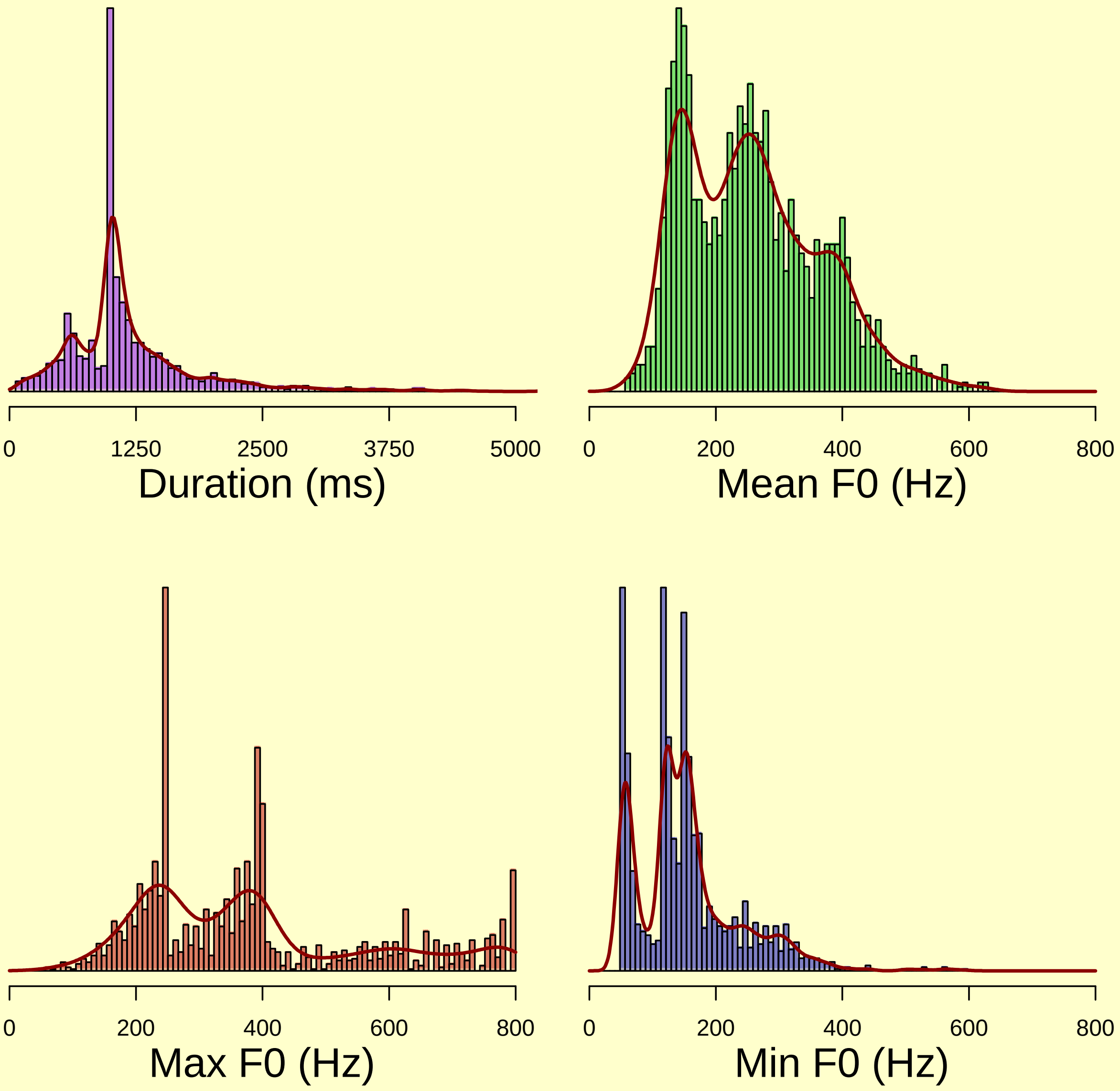


Reliability of LENA labels, current findings

How do acoustic factors drive human and machine classification decisions for families with HH kid?

FACTORS:

1. duration
2. *f*0 – mean
3. *f*0 – min
4. *f*0 – max
5. *f*0 – rise
6. *f*0 – fall
7. amp, RMS
8. amp, rise
9. amp, fall



Conclusions

1. Human and machine decisions use a <u>similar subset of factors</u> in decision making; duration and *f*0 appear to be important, but amplitude does not appear to play a role.

2. Machines and humans use <u>similar strategies</u> to assign talker labels to acoustic input:
***f*0 dominates duration, and amplitude is not too important.**

3. Machine and humans seem to treat TD and HH data similarly; very long duration (>970ms) may be unique to TD kids; interestingly, some of the *f*0 or *f*0-contour did not seem to be unique to TD kids.

4. Data is <u>messy</u>. Individual difference, algorithm artifacts (whisper, singing, range parameters), may influence machine output, but we can only speculate.

5. Other factors may play a role: spectral envelope/mean/tilt, shimmer (amp error), jitter (*f*0 error), SNR, nasalance, vocal quality (creak, fry), etc.