A Longitudinal Study of Infant Speech Production Parameters: A Case Study

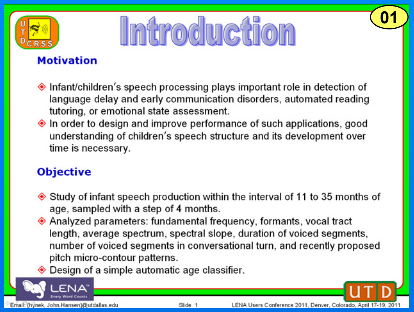
Hynek Bořil[1], John H.L. Hansen[1], Dongxin Xu[2], Gill Gilkerson[2], Jeff Richards[2]

{hynek, John.Hansen}@utdallas.edu

[1]Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

[2]LENA Foundation
Boulder, Colorado, USA

LENA Users Conference 2011: April 17-19, 2011 Denver, Colorado, USA



01

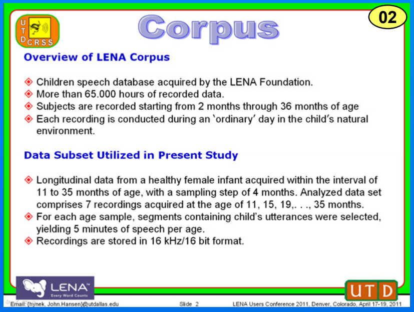
Introduction

Motivation

- Infant/children's speech processing plays important role in detection of language delay and early communication disorders, automated reading tutoring, or emotional state assessment.
- In order to design and improve performance of such applications, good understanding of children's speech structure and its development over time is necessary.

Objective

- Study of infant speech production within the interval of 11 to 35 months of age, sampled with a step of 4 months.
- Analyzed parameters: fundamental frequency, formants, vocal tract length, average spectrum, spectral slope, duration of voiced segments, number of voiced segments in conversational turn, and recently proposed pitch micro-contour patterns.
- Design of a simple automatic age classifier.



02

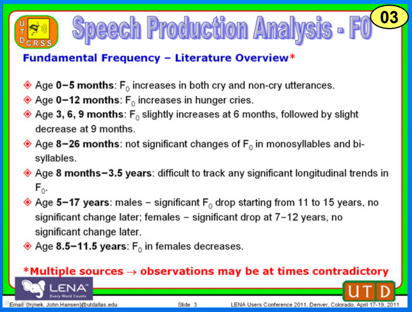
Corpus

Overview of LENA Corpus

- Children speech database acquired by the LENA Foundation.
- More than 65.000 hours of recorded data.
- Subjects are recorded starting from 2 months through 36 months of age
- Each recording is conducted during an 'ordinary' day in the child's natural environment.

Data Subset Utilized in Present Study

- Longitudinal data from a healthy female infant acquired within the interval of 11 to 35 months of age, with a sampling step of 4 months. Analyzed data set comprises 7 recordings acquired at the age of 11, 15, 19,. . ., 35 months.
- For each age sample, segments containing child's utterances were selected, yielding 5 minutes of speech per age.
- Recordings are stored in 16 kHz/16 bit format.



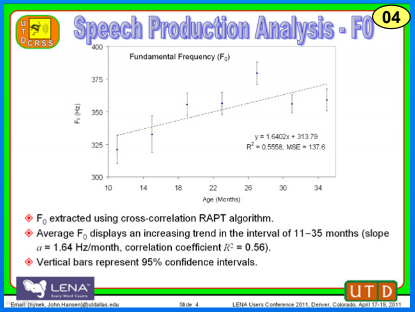
03

Speech Production Analysis - F0

Fundamental Frequency – Literature Overview*

- Age **0–5 months**: F_0 increases in both cry and non-cry utterances.
- Age **0–12 months**: F_0 increases in hunger cries.
- Age **3, 6, 9 months**: F_0 slightly increases at 6 months, followed by slight decrease at 9 months.
- Age **8–26 months**: not significant changes of F_0 in monosyllables and bi-syllables.
- Age **8 months–3.5 years**: difficult to track any significant longitudinal trends in F_0.
- Age **5–17 years**: males – significant F_0 drop starting from 11 to 15 years, no significant change later; females – significant drop at 7–12 years, no significant change later.
- Age **8.5–11.5 years**: F_0 in females decreases.

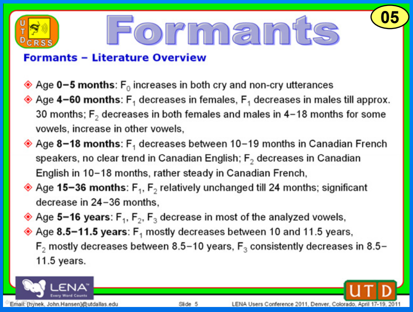
*Multiple sources → observations may be at times contradictory



04

Speech Production Analysis - F0

- F_0 extracted using cross-correlation RAPT algorithm.
- Average F_0 displays an increasing trend in the interval of 11–35 months (slope a = 1.64 Hz/month, correlation coefficient R^2 = 0.56).
- Vertical bars represent 95% confidence intervals.

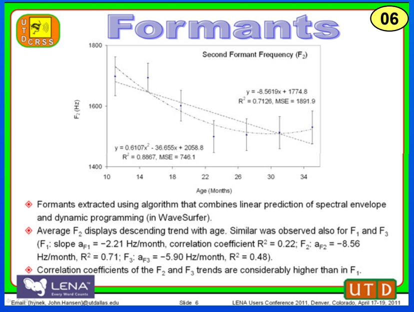


05

Formants

Formants – Literature Overview

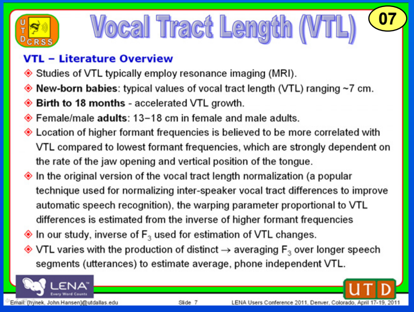
- Age **0–5 months**: F_0 increases in both cry and non-cry utterances
- Age **4–60 months**: F_1 decreases in females, F_1 decreases in males till approx. 30 months; F_2 decreases in both females and males in 4–18 months for some vowels, increase in other vowels,
- Age **8–18 months**: F_1 decreases between 10–19 months in Canadian French speakers, no clear trend in Canadian English; F_2 decreases in Canadian English in 10–18 months, rather steady in Canadian French,
- Age **15–36 months**: F_1, F_2 relatively unchanged till 24 months; significant decrease in 24–36 months,
- Age **5–16 years**: F_1, F_2, F_3 decrease in most of the analyzed vowels,
- Age **8.5–11.5 years**: F_1 mostly decreases between 10 and 11.5 years, F_2 mostly decreases between 8.5–10 years, F_3 consistently decreases in 8.5–11.5 years.



06

Formants

- Formants extracted using algorithm that combines linear prediction of spectral envelope and dynamic programming (in WaveSurfer).
- Average F_2 displays descending trend with age. Similar was observed also for F_1 and F_3 (F_1: slope $a_{F1} = -2.21$ Hz/month, correlation coefficient $R^2 = 0.22$; F_2: $a_{F2} = -8.56$ Hz/month, $R^2 = 0.71$; F_3: $a_{F3} = -5.90$ Hz/month, $R^2 = 0.48$).
- Correlation coefficients of the F_2 and F_3 trends are considerably higher than in F_1.

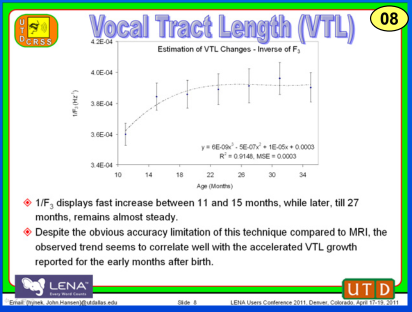


07

Vocal Tract Length (VTL)

VTL – Literature Overview

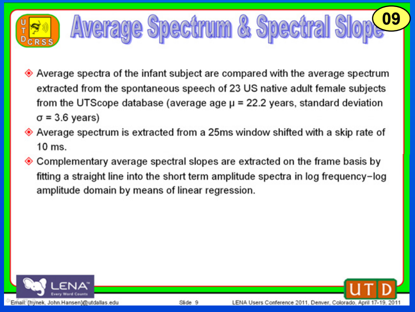
- Studies of VTL typically employ resonance imaging (MRI).
- **New-born babies**: typical values of vocal tract length (VTL) ranging ~7 cm.
- **Birth to 18 months** - accelerated VTL growth.
- Female/male **adults**: 13–18 cm in female and male adults.
- Location of higher formant frequencies is believed to be more correlated with VTL compared to lowest formant frequencies, which are strongly dependent on the rate of the jaw opening and vertical position of the tongue.
- In the original version of the vocal tract length normalization (a popular technique used for normalizing inter-speaker vocal tract differences to improve automatic speech recognition), the warping parameter proportional to VTL differences is estimated from the inverse of higher formant frequencies
- In our study, inverse of F_3 used for estimation of VTL changes.
- VTL varies with the production of distinct → averaging F_3 over longer speech segments (utterances) to estimate average, phone independent VTL.



08

Vocal Tract Length (VTL)

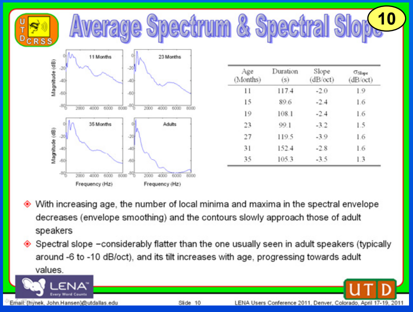
- $1/F_3$ displays fast increase between 11 and 15 months, while later, till 27 months, remains almost steady.
- Despite the obvious accuracy limitation of this technique compared to MRI, the observed trend seems to correlate well with the accelerated VTL growth reported for the early months after birth.



09

Average Spectrum & Spectral Slope

- Average spectra of the infant subject are compared with the average spectrum extracted from the spontaneous speech of 23 US native adult female subjects from the UTScope database (average age μ = 22.2 years, standard deviation σ = 3.6 years)
- Average spectrum is extracted from a 25ms window shifted with a skip rate of 10 ms.
- Complementary average spectral slopes are extracted on the frame basis by fitting a straight line into the short term amplitude spectra in log frequency–log amplitude domain by means of linear regression.

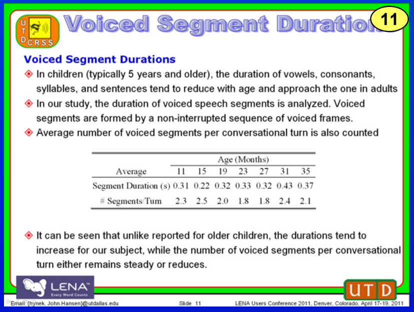


10

Average Spectrum & Spectral Slope

Age (Months)	Duration (s)	Slope (dB/oct)	σ_{Slope} (dB/oct)
11	117.4	-2.0	1.9
15	89.6	-2.4	1.6
19	108.1	-2.4	1.6
23	99.1	-3.2	1.5
27	119.5	-3.9	1.6
31	152.4	-2.8	1.6
35	105.3	-3.5	1.3

- With increasing age, the number of local minima and maxima in the spectral envelope decreases (envelope smoothing) and the contours slowly approach those of adult speakers
- Spectral slope –considerably flatter than the one usually seen in adult speakers (typically around -6 to -10 dB/oct), and its tilt increases with age, progressing towards adult values.



11

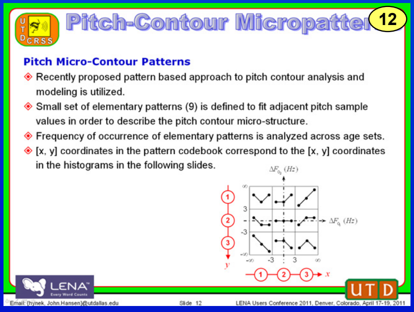
Voiced Segment Duration

Voiced Segment Durations

- In children (typically 5 years and older), the duration of vowels, consonants, syllables, and sentences tend to reduce with age and approach the one in adults
- In our study, the duration of voiced speech segments is analyzed. Voiced segments are formed by a non-interrupted sequence of voiced frames.
- Average number of voiced segments per conversational turn is also counted

	Age (Months)						
Average	11	15	19	23	27	31	35
Segment Duration (s)	0.31	0.22	0.32	0.33	0.32	0.43	0.37
# Segments/Turn	2.3	2.5	2.0	1.8	1.8	2.4	2.1

- It can be seen that unlike reported for older children, the durations tend to increase for our subject, while the number of voiced segments per conversational turn either remains steady or reduces.

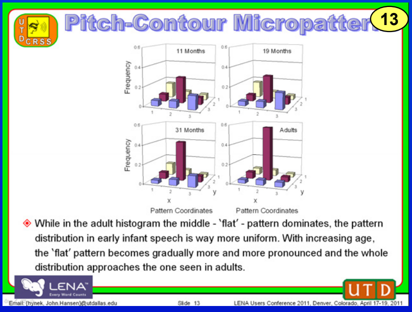


12

Pitch-Contour Micropatter

Pitch Micro-Contour Patterns

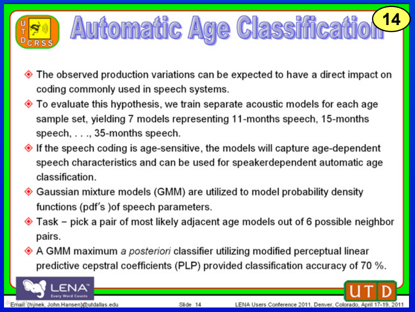
- Recently proposed pattern based approach to pitch contour analysis and modeling is utilized.
- Small set of elementary patterns (9) is defined to fit adjacent pitch sample values in order to describe the pitch contour micro-structure.
- Frequency of occurrence of elementary patterns is analyzed across age sets.
- [x, y] coordinates in the pattern codebook correspond to the [x, y] coordinates in the histograms in the following slides.



13

Pitch-Contour Micropatter

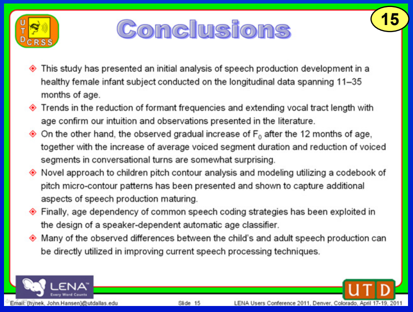
- While in the adult histogram the middle - 'flat' - pattern dominates, the pattern distribution in early infant speech is way more uniform. With increasing age, the 'flat' pattern becomes gradually more and more pronounced and the whole distribution approaches the one seen in adults.



14

Automatic Age Classification

- The observed production variations can be expected to have a direct impact on coding commonly used in speech systems.
- To evaluate this hypothesis, we train separate acoustic models for each age sample set, yielding 7 models representing 11-months speech, 15-months speech, . . ., 35-months speech.
- If the speech coding is age-sensitive, the models will capture age-dependent speech characteristics and can be used for speakerdependent automatic age classification.
- Gaussian mixture models (GMM) are utilized to model probability density functions (pdf's)of speech parameters.
- Task – pick a pair of most likely adjacent age models out of 6 possible neighbor pairs.
- A GMM maximum *a posteriori* classifier utilizing modified perceptual linear predictive cepstral coefficients (PLP) provided classification accuracy of 70 %.



15

Conclusions

- This study has presented an initial analysis of speech production development in a healthy female infant subject conducted on the longitudinal data spanning 11–35 months of age.
- Trends in the reduction of formant frequencies and extending vocal tract length with age confirm our intuition and observations presented in the literature.
- On the other hand, the observed gradual increase of F_0 after the 12 months of age, together with the increase of average voiced segment duration and reduction of voiced segments in conversational turns are somewhat surprising.
- Novel approach to children pitch contour analysis and modeling utilizing a codebook of pitch micro-contour patterns has been presented and shown to capture additional aspects of speech production maturing.
- Finally, age dependency of common speech coding strategies has been exploited in the design of a speaker-dependent automatic age classifier.
- Many of the observed differences between the child's and adult speech production can be directly utilized in improving current speech processing techniques.