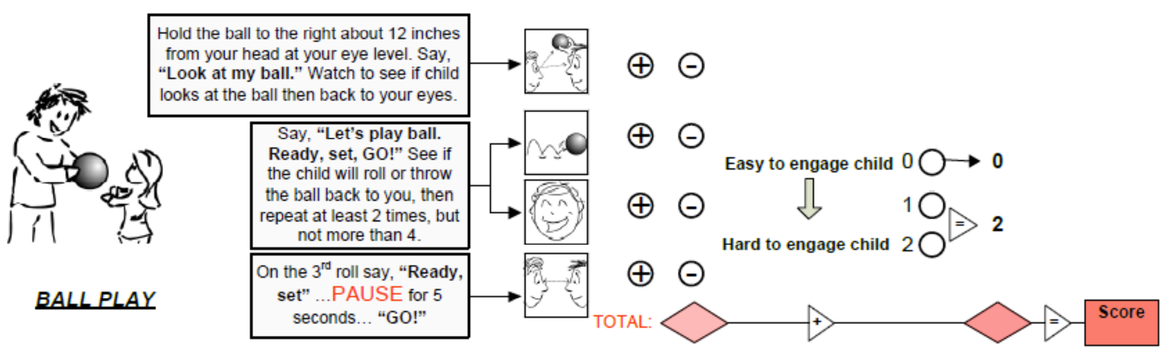
ENGAGEMENT LEVEL ESTIMATION OF SEMI-STRUCTURED CHILD-EXAMINER INTERACTIONS FROM SPEECH EVENTS

Jonathan Kim, Hrishikesh Rao, and Mark A. Clements
Georgia Institute of Technology

Expeditions in Computing

Introduction

Rapid ABC consists of five activities: greeting, ball play, book reading, hat play, and tickling. Analysis of the speech from both the examiner and the child over 46 Rapid ABC sessions suggests it is possible to predict the level of engagement in all five activities. Since the vast majority of scores were 0 (easy to engage), the three levels of engagement were grouped into two classes, {0} and {1, 2}, to achieve better balance.

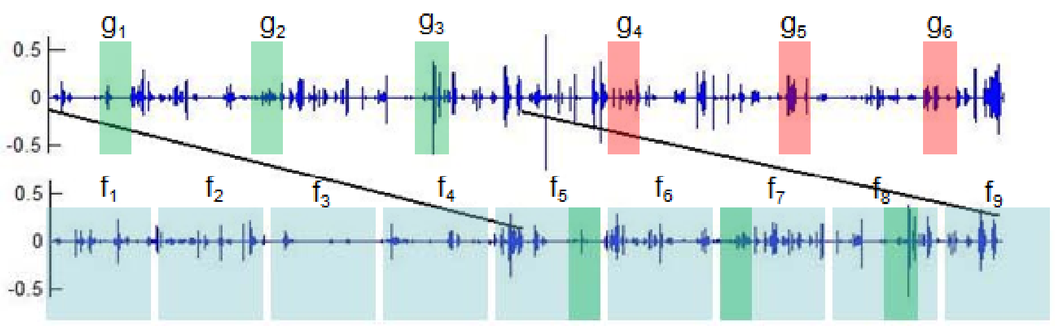


Approach

1.1 Audio Synchronization

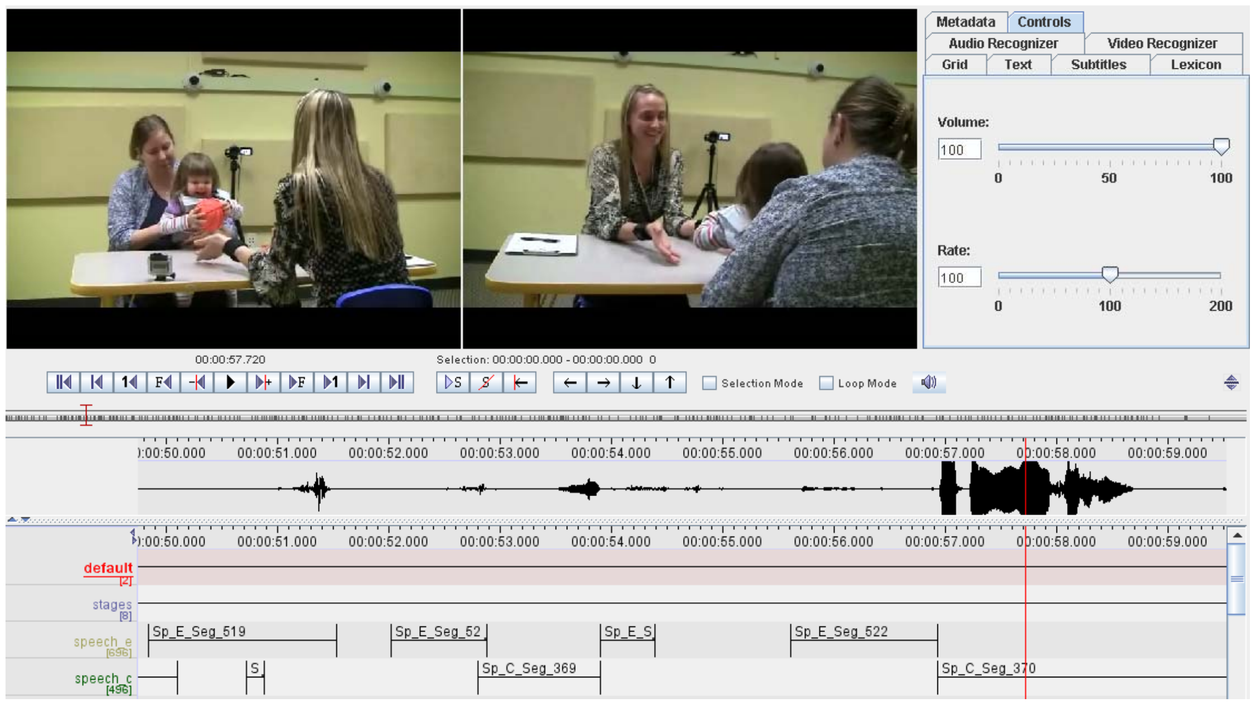
The audio data are synchronized by finding the maximum values of the cross-correlation functions of audio signals. The multiple local maxima are found by dividing the two original signals into various length chunks before computing cross-correlations.

$$c[i,k] = \left(\arg\max_{0 \le n \le M} \sum_{m=0}^{L} | f_k[m] g_i[n+m] | \right) + (k-1)M - (i-1)D$$



1.2 Speech Segmentation into Discrete Phrases

a) Energy threshold
b) SNR calculation with noise adaptation
c) Zero-crossing rates
d) Pitch estimation
e) Voiced/unvoiced ratio



2.1 Acoustic Feature Extraction

For each analysis frame, energy, pitch, harmonic-to-noise ratio, mel-frequency cepstral coefficients (MFCCs) and their derivatives were measured. For segment level feature extraction, the time series of all the measurements were represented with the statistical and regressional measures.

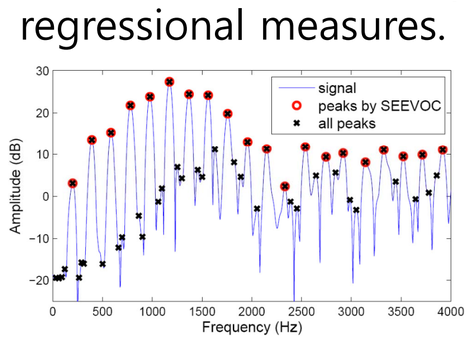


Fig. 1: Harmonic peaks selected by SEEVOC picking routine.

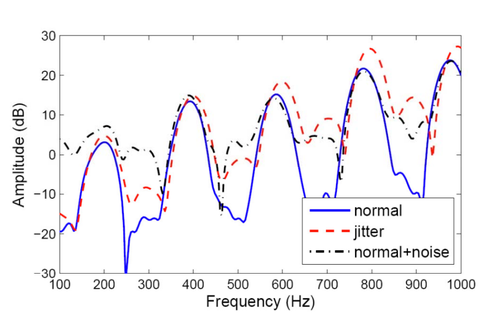


Fig. 2: Comparison of normal, pitch jitter, and normal+noise signals in a low frequency band.

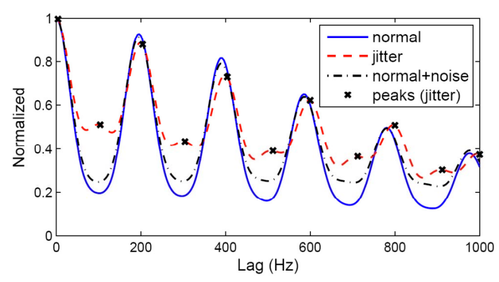


Fig. 3: R(f) of normal, pitch jitter, and normal+noise signals.

$$HIR = \frac{L}{K} \frac{\sum_{k=1}^{K} A_k}{\sum_{l=1}^{L} P_l}$$

$$Mean(\frac{\Delta f}{f_0}) = \frac{1}{K-1} \sum_{k=1}^{K-1} \frac{(f_{k+1} - f_k)}{f_0}$$

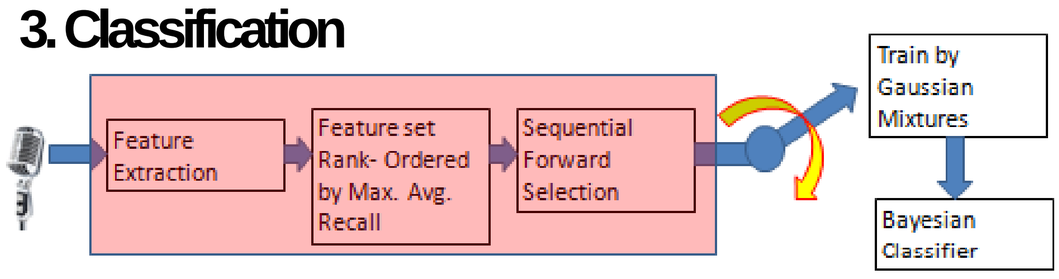
Table 1: Low-level descriptors (LLDs)

Type	Measure
Statistical	Maximum, minimum, mean, standard deviation, kurtosis, skewness, flatness1^{st}, 2^{nd} & 3^{rd} quartiles, interquartile range, 1^{st} & 99^{th} percentiles, and root mean square value
Regressional	Slope of linear regression, approximation error of linear regression, quadratic regression coefficient, and approximation error of quadratic regression

2.2 Event-based Features

1	Duration of Stage	12	Num C-to-E / Num Examiner Segs
2	Speech Duration of Child	13	Num C-to-E / Num Child Segs
3	Speech Duration of Examiner	14	Num C-to-E / (Num Examiner + Child Segs)
4	Duration of X-Talk	15	Num E-to-C / Num Child Segs
5	Number of C-to-E	16	Num E-to-C / Num Examiner Segs
6	Number of E-to-C	17	Num E-to-C / (Num Examiner + Child Segs)
7	Number of Child Speech Segments	18	Avg Child Speech Seg Dur
8	Number of Examiner Speech Segments	19	Avg Examiner Speech Seg Dur
9	Dur Child / Dur Stage	20	Max Child Speech Seg Dur
10	Dur Exam / Dur Stage	21	Max Examiner Speech Seg Dur
11	Dur X-talk / Dur Stages		

3. Classification



Results

Number of RABC sessions analyzed : 46
Number of participants: 31
Duration of 46 sessions: 152 minutes

Greet	H′(0)	H′(1)
H(0)	53.3	46.7
H(1)	45.0	54.9

UWA = 54.11 %
WA = 53.78 %

Ball	H′(0)	H′(1)
H(0)	85.3	14.6
H(1)	17.9	82.0

UWA = 83.72 %
WA = 84.49 %

Book	H′(0)	H′(1)
H(0)	74.9	25.0
H(1)	45.5	54.4

UWA = 64.69 %
WA = 63.31 %

Hat	H′(0)	H′(1)
H(0)	91.6	8.33
H(1)	70.5	29.5

UWA = 60.58 %
WA = 86.34 %

Tickle	H′(0)	H′(1)
H(0)	80.9	19.0
H(1)	47.9	52.0

UWA = 66.51 %
WA = 73.41 %