DETECTION OF LAUGHTER IN CHILDREN'S SPEECH USING SPECTRAL AND PROSODIC ACOUSTIC FEATURES

Hrishikesh Rao[1], Jonathan C. Kim[1], Agata Rozga[2], and Mark A. Clements[1]

[1] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2] School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

Expeditions in Computing

Motivation

- Laughter is an important para-linguisitc cue that can be useful in gauging the affective state of the speaker.
- Detection of laughter in children's speech is less explored and has important applications in clinical psychology.
- Laughter , along with other vocalizations, is an important marker for very early detection of autism spectrum disorder (ASD) [1].
- Diarization of para-lingusitic events would benefit psychologists who are interested in studying children's affective communication.

Datasets

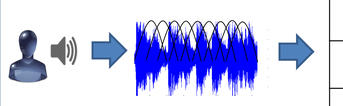
FAU-Aibo Emotion Corpus [2], [3]

- Interaction between adolescent and Sony's Aibo robot
- Data collected from 51 children (aged 10-13 years, 21 male, 30 female)
- Laughter annotated along with speech.
- Different types of laughter annotated (voiced, unvoiced, voiced-unvoiced, and speech modulated with laughter)
- Number of sample points – 13478 for speech and 236 for laughter.

Rapid ABC Dataset [4]

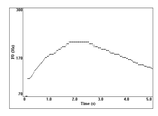
- Semi-structured dyadic interaction between toddler and examiner
- Activities include greeting the child, initiating a game of rolling the ball back and forth, bringing a book and inviting the child to through it, pretending the book to be a hat, and engaging the child in a tickling game.
- 20 Rapid ABC sessions (aged 15-29 months) used with laughter and other vocalizations annotated.
- Number of sample points – 17 each for speech and laughter.

Methodology and Results

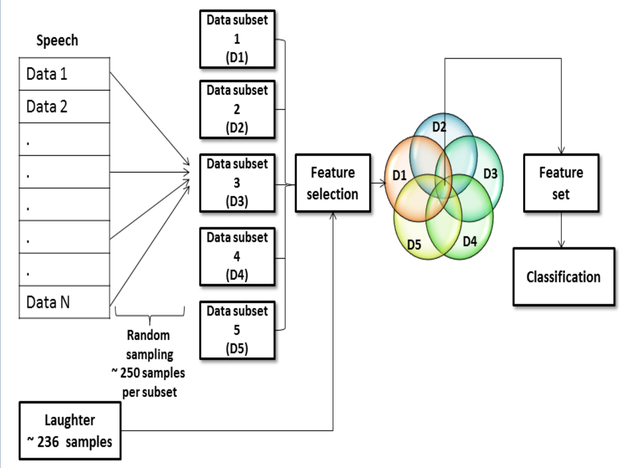


Hamming window of 30 ms with 10 ms overlap

Feature	Number of low-level descriptors
Intensity	2
Loudness	2
Mel-frequency Cepstral Coefficients	24
Pitch	2
Probability of voicing	2
Pitch envelope	2
Line spectral frequencies	16
Zero-crossing rate	2
Formant frequency and bandwidth	6
Ratio of formant frequencies	3
Ratio of bandwidths of formants	3
Euclidean distance between formant frequencies	3
Euclidean distance between formant bandwidths	3
Euclidean distance between ratio of formant frequencies	3



Statistical Measure
Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges



- Five subsets of data consisting of randomly selected 250 speech samples
- Frame-level acoustic features extracted and statistical measures evaluated at the phrase level
- Features are ranked as per the information gain criterion
- $IG(w_i, Xj) = -\sum_{i=1}^{M} \Pr(w_i)\, log_2 \Pr(wi) + \sum_{j=1}^{N}\sum_{i=1}^{M} \Pr(X_j)\, \Pr(w_i|X_j)\, log_2 \Pr(wi|Xj)$
- Intersection of top 100 features for each subset results in final feature set

Feature	Number of features selected
Probability of voicing	12
Pitch	5
Mel-frequency Cepstral Coefficient	5
Line Spectral Frequency	3
First Formant Frequency	5

Task: 10-fold cross-validation using five subsets of data with a various classifiers

Classifier	Accuracy (mean ± standard deviation)
Multi-layer Perceptron	95.04±2.67%
Radial Basis Function Neural Networks	95.44±2.70%
SVM (Linear kernel)	95.30±2.68%
SVM (Polynomial kernel, degree = 2)	95.82±2.27%
SVM (RBF kernel)	95.96±2.28%
GMM - EM	95.16±3.25%

Task: Clustering with GMM-EM and k-Means using five subsets of data

Clustering algorithm	Error rate (mean ± standard deviation)
k-Means	7.19±3.67%
GMM-EM	5.71±3.16%

Clustering results indicate robust predictive power of selected features

Task: Classification using a support vector machine (SVM) with a polynomial kernel (degree = 1.65) on FAU-AEC dataset

	Predicted Speech	Predicted Laughter
True Speech	12726	752
True Laughter	13	223

Weighted accuracy : **94.43%**
Unweighted accuracy : **94.46%**

FAU-AEC's Results
Weighted accuracy : **81.95%**
Unweighted accuracy : **82.95%**

Absolute improvement of **12.48%** over FAU-AEC

Task: Classification using a support vector machine (SVM) with a linear kernel trained using FAU-AEC dataset and tested on Rapid ABC dataset

	Predicted Speech	Predicted Laughter
True Speech	12	5
True Laughter	5	12

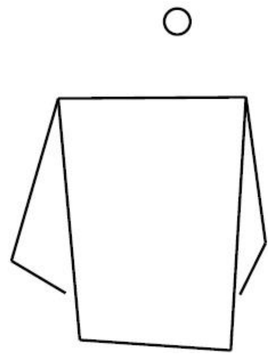
Weighted accuracy : **70.58%**
Unweighted accuracy : **70.58%**

Multi-Modal Analysis Ideas

Laughter and Smile Detector

- Laughter correlated with smiling.
- FaceTracker used for extracting visual features.
- Late fusion of scores from visual and audio classifier could be used for predicting emotion viz. joy.

Posture and Para-Lingusitic Event Detector



- Tickling involves movement of upper body and laughter.
- Use of upper body predicates along with laughter detection results.
- Fusion of visual and audio features for predicting level of engagement.
- Could be used to parse the tickling state of Rapid ABC.

Conclusions

- Robust detection of laughter is possible in children's speech using acoustic features
- Statistically relevant generalization on other datasets consisting of different recording conditions, ages of subjects, and languages.
- Multimodal analysis using vision and electro-dermal activity improves the understanding of the affective nature of laughter.

Acknowledgment

The authors would like to thank the National Science Foundation (NSF) for their gracious support towards this research (NSF grant No. CCF-1029679)