Classifying Language-Related Developmental Disorders from Speech Cues: the Promise and the Potential Confounds

Daniel Bone, **Theodora Chaspari, Kartik Audkhasi, James Gibson, Andreas Tsiartas, Maarten Van Segbroeck, Ming Li, Sungbok Lee, Shrikanth Narayanan**

Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA

presented by Ladislav Rampasek

Outline

- Interspeech 2013 Autism Sub-Challenge
  - **4 groups of children speakers**
- study of features that may inform realistic separability between groups
- potential confounds in the data

INTERSPEECH 2013
AUTISM SUB-CHALLENGE

Goal

- Determine the type of pathology of a speaker:
  - autism spectrum disorders (**ASD**)
  - specific language impairment (**SLI**)
  - pervasive developmental disorder - not otherwise specified (**PDD-NOS**)
  - typically developing (**TD**)

…from short audio recordings

autism spectrum disorders (**ASD**)

- Includes:
  - autistic disorders
  - Asperger's disorders
  - and newly also PDD-NOS
- impaired social communication
- restricted, repetitive, and/or stereotyped behavioral patterns
- **impaired receptive and expressive prosody,** but no established prevalence estimates of subjective prosodic abnormalities

specific language impairment (**SLI**)

- developmental dysphasia or developmental aphasia
- **speech prosody has been understudied** (because seen as unlikely)
- however some evidence does suggest **impaired reception and production of prosody**

Data

- 2542 instances of speech recordings from 99 children aged 6 to 18
- by 2 university departments of child and adolescent psychiatry, in Paris, France

#	train	dev	test	Σ
Typically developing				
TYP	566	543	542	1651
Atypically developing				
ASD	104	104	99	307
PDD-NOS	104	68	75	247
SLI	129	104	104	337
Σ	903	819	820	2542

Audio Recordings

- French-speaking participants
- **intonation imitation task:** attempting to accurately reproduce perceived lexical and prosodic information
- ranging from 170 ms to 7.2 s (mean = 1.4 s)
- prompted 26 sentences representing
  - 4 different *modalities*: **declarative**, **exclamatory**, **interrogative**, and **imperative**
  - 4 types of *intonations*: **descending**, **falling**, **floating**, and **rising**

Baseline

- 6,373 features from openSMILE e.g.:
  - energy, spectral, cepstral (MFCC) and voicing related low-level descriptors
  - logarithmic harmonic-to-noise ratio, spectral harmonicity, and psychoacoustic spectral sharpness
- model:
  - SVM, and synthetic sampling to balance classes

Two Classification Tasks

1. binary **Typicality** task:
   - typically vs. atypically developing children
   - baseline = 92.8% unweight average recall

2. four-way **Diagnosis** task:
   - classifying into ASD, SLI, PDD-NOS, TD
   - baseline = 51.7% unweight average recall

THIS PAPER

Main Focus

1. study of features that may inform realistic separability between groups
   - **prosodic and formant templates**
   - **pronunciation quality**

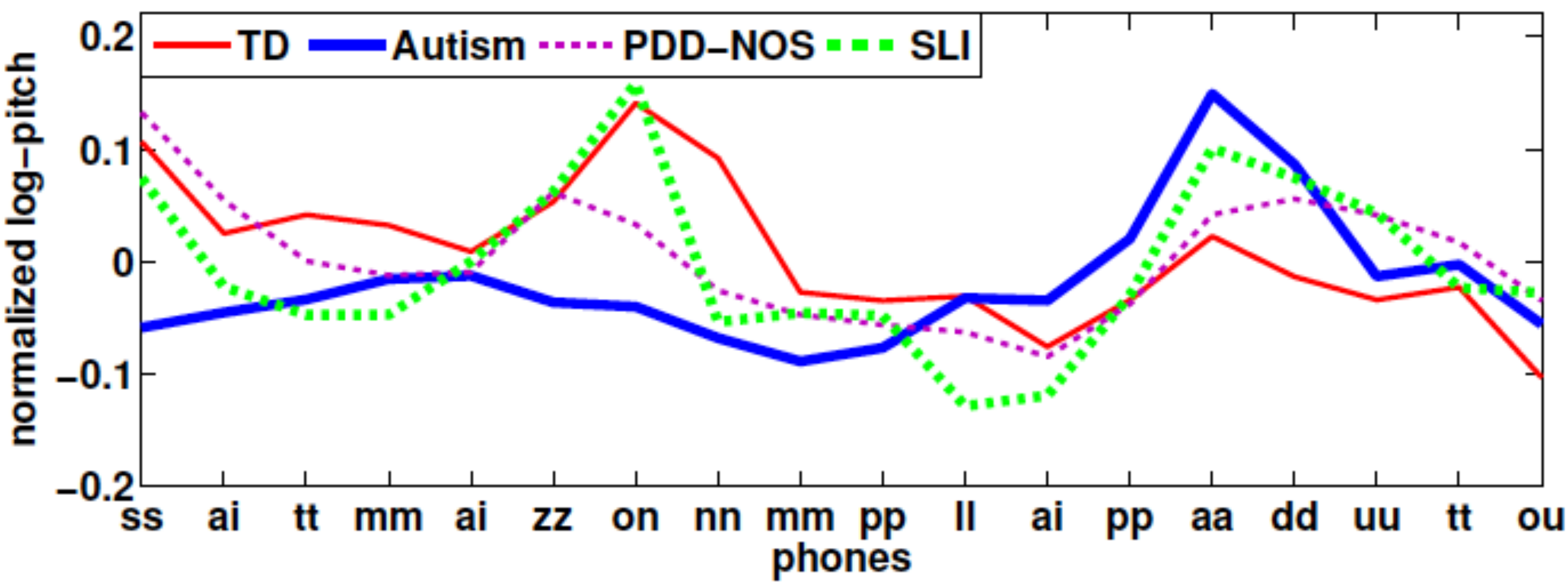
2. potential confounds in the data
   - the baseline, and spectral-based methods are most likely over-fitting to the channel effects (like reverberation)

Prosodic and Formant Templates

- contour templates constructed across phones (using forced-alignments):
  - **pitch** contour templates
  - **intensity** contour templates
  - **duration** contour templates
  - and **formant** contour templates
- optimal reproduction templates:
  - generated from the typically developing speakers recordings in the training data

Normalized log-pitch contours

"Cette maison ne me plait pas du tout."
"This house does not please me at all."



Contours Computation

- constructed across phones (each consecutive phone represents a point in time)
- features computed within the boundaries of a phone
- for **log-pitch**, **formants** (F1-F3), and **intensity**:
  - modeled as a 2nd order polynomial
  - ⇒3 contours per feature (corresponding to *curvature*, *slope*, and *zero-crossing*)
- the **duration** contour is simply the duration of each phone

Templates Computation

- computed per sentence as the median feature value for each phone
- using only utterances from typical development speakers

- 2 features between **template** and **contour**:
  1. Correlation
  2. Mean absolute difference (L1 norm)

Pronunciation Quality

- The **goodness of pronunciation** (GOP) score:
  - average log-posterior probability of each reference phone *p* from the output of an ASR system:

$$GOP(p) = -\log P(p \mid \mathbf{o}^{p}) / NF(p)$$

  - $\mathbf{o}^{p}$ = acoustic observation sequence for phone *p*
  - NF(*p*) = corresponding number of frames

The Model

for prosodic-template and goodness of pronunciation features

- linear-kernel SVM model
- these features require the utterance to be known
- thus utterance recognition (ASR) was developed on the development set

Results (robust features)

	2-class	4-class
Chance	50	25
Development Set Baseline	92.8	51.7
Total Duration (Per-Sentence)	61.4	29.6
Pitch Template (P)	64.1	32.0
Duration Template (D); ***P+D***	69.9; *73.4*	39.5; *38.0*
Formants Template (F); ***P+D+F***	62.4; *74.3*	34.4; *33.7*
Intensity Template (I); ***P+D+F+I***	70.2; *79.7*	34.9; *38.2*
Goodness of Pron. (Per-Sentence)	68.1	29.9
Spectral Energy and Smoothness	92.7	62.4

Spectral Energy and Smoothness

- 360 features that capture spectrogram energy levels and variations
  - e.g. total signal energy, mean and relative energy changes over multiple time scales and frequency bands, and the frequencies with the majority of energy content
- + long-term functionals of these features
- + MFCC and RASTA-PLP features
- = total of 386 features

The Model
for frame-level spectral energy features

- forward feature selection
- k-NN classifier

- 5 features for the 2-class task selected
- 7 features for the 4-class task selected
- *unclear how much, these spectral variations actually are due to the differences in the health conditions* => **picking up channel effects?**

Results (spectral features)

	2-class	4-class
Chance	50	25
Development Set Baseline	92.8	51.7
Total Duration (Per-Sentence)	61.4	29.6
Pitch Template (P)	64.1	32.0
Duration Template (D); *P+D*	69.9; *73.4*	39.5; *38.0*
Formants Template (F); *P+D+F*	62.4; *74.3*	34.4; *33.7*
Intensity Template (I); *P+D+F+I*	70.2; *79.7*	34.9; *38.2*
Goodness of Pron. (Per-Sentence)	68.1	29.9
Spectral Energy and Smoothness	92.7	62.4

Ensemble of Models

- 2 models linear-kernel SVMs with baseline features
- 2 deep neural networks with baseline features
- 1 model based on spectral energy features with k-NN classification

Ensemble of Models

- SMOTE up-sampling and hierarchical classification structure:
  - Typical vs. Atypical
  - ASD vs. SLI
  - PDD-NOS vs. ASD

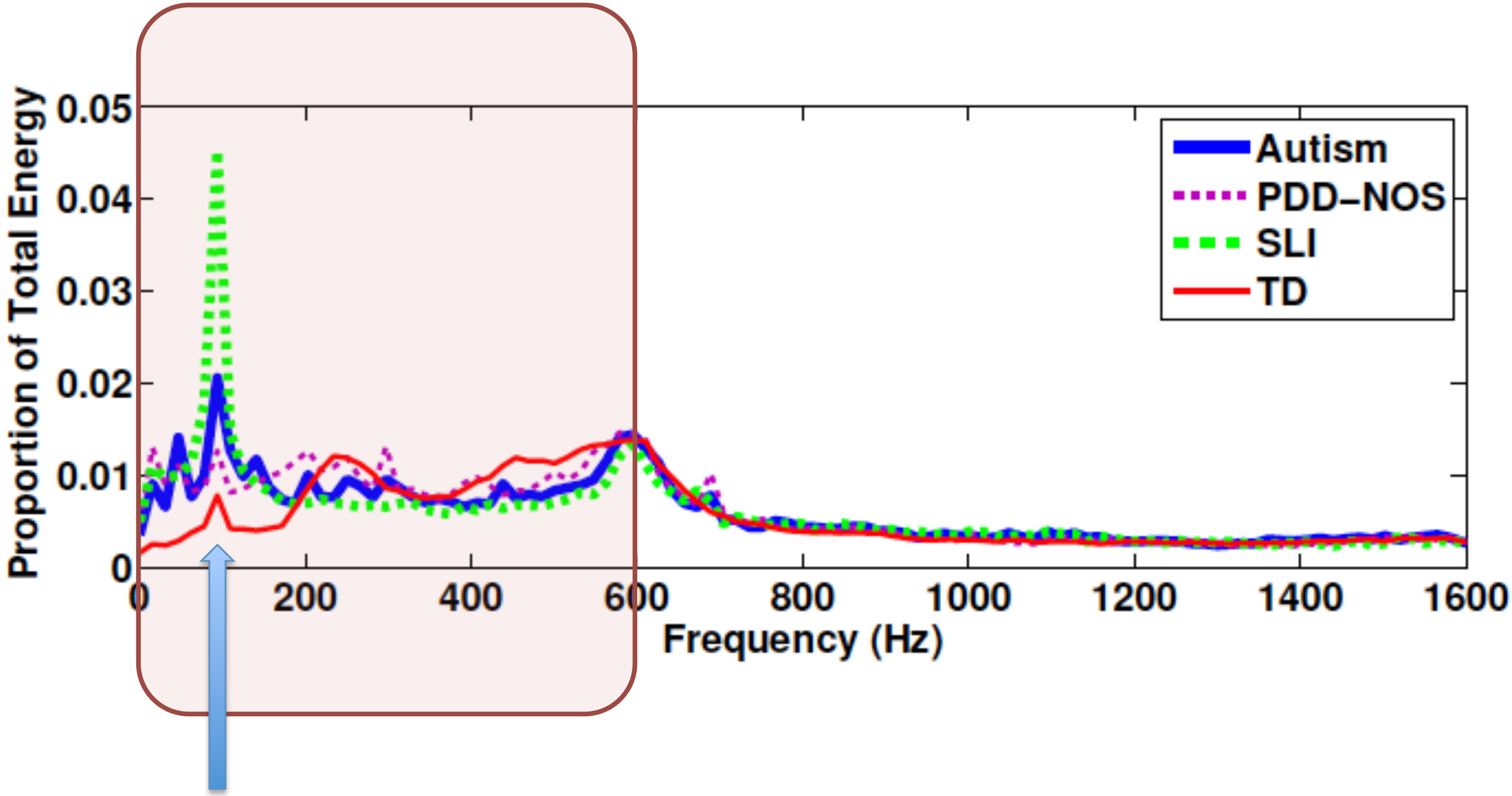
- achieved accuracy of **60.2**% UAR

Variability in Acoustic Environments: Effect on Signal Features

- authors noticed distinct reverberation in the typically developing data compared to the language impaired data recordings
- from the short recordings it's difficult to quantify such room acoustic properties
- instead they looked at differences in the long-term average spectrum of the recordings

Mean Normalized *Long Term Average Spectrum*

Differences between groups appear below 600 Hz, mainly below 400Hz



spikes of varying height near 100 Hz, possibly an electric hum harmonic

Classification by Single Gaussian

1. trained on the LTAS of audio recordings from each group
2. then, maximum-likelihood decisions for each utterance in the development set

- using normalized energy bins of 0-400 Hz, they got:
  - **79.7**% 2-way (below baseline)
  - **51.4**% 4-way (ties baseline)

Effect on Signal Features

- long-term spectral characteristics could reflect room acoustics and voice quality characteristics, as opposed to lexical content, especially as all groups spoke the same utterance
- the precise cause and scope of channel effects is hard to estimate from such short recordings
- authors conclude variations in recording environments do exist and influence the results

Conclusion

- achieved above chance accuracies by using **prosodic template** and **pronunciation quality** modeling
- these features are **likely to generalize well**
- **comb** … appears most …
- resul … most differ …
- but n … o is generally as different from the **TD** group
- surprisingly high accuracy of the spectral-energy methods **suggest significant channel effects**

"Therefore, the performance differences between populations **are unclear** from our study."