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

http://sail.usc.edu

Abstract

Speech and spoken language cues offer a valuable means to measure and model human behavior. Computational models of speech behavior have the potential to support health care through assistive technologies, informed intervention, and efficient long-term monitoring. The Interspeech 2013 Autism Sub-Challenge addresses two developmental disorders that manifest in speech: autism spectrum disorders and specific language impairment. We present classification results with an analysis on the development set including a discussion of potential confounds in the data such as recording condition differences. We hence propose study of features within these domains that may inform realistic separability between groups as well as have the potential to be used for behavioral intervention and monitoring. We investigate template-based prosodic and formant modeling as well as goodness of pronunciation modeling, reporting above chance classification accuracies.

Index Terms: autism spectrum disorders, intonation, specific language impairment, goodness of pronunciation

1. Introduction

Observational analysis of speech and spoken language is a central facet of diagnostics and intervention in the behavior sciences. Behavioral signal processing (BSP) aims to use computational methods to inform the assessment of human behavior, with techniques potentially being applied to inform interventions and to advanced human-machine interfaces. The three components of BSP are: (1) acquisition of ecologically-valid behavioral signal data; (2) analysis to create behavioral descriptors; and (3) modeling of the mapping between behavioral descriptors and behavioral constructs [1].

Autism spectrum disorders (ASD), also known as autism spectrum conditions (ASC), are developmental disorders that result in impaired social communication and reciprocity, as well as restricted, repetitive, and/or stereotyped behavioral patterns [2]. Social-affective impairments in ASD often present in the form of impaired receptive and expressive prosody [3]. Specific language impairment, or SLI (historically also referred to as developmental dyphasia or developmental aphasia), is a developmental language disorder that occurs in the absensce of co-morbid conditions such as hearing loss, neurological trauma, or low non-verbal IQ [4]. SLI diagnosis is conducted using standardized language tests that index phonology, vocabulary, and syntax. Links between ASD and SLI have been proposed, with a population lying in the spectrum between the traditional definitions of the two disorders [5].

Signal processing and machine learning afford potential advances in supporting characterization and treatment of developmental disorders like ASD and SLI. In autism, qualitative descriptions of atypical prosody are widespread [6], but prevalence estimates of subjective prosodic abnormalities as well as established objective measures are lacking. In SLI, speech prosody has been understudied due to a view that intonation is unlikely to be afffected by developmental speech and language impairment. However, others have suggested that short-term memory deficits in SLI will lead to prosodic difficulties; additionally, some evidence does suggest impaired reception and production of prosody in SLI [7].

The Autism Sub-Challenge of the Interspeech 2013 Computational Paralinguistic Challenge asks participants to determine the type of pathology of a speaker (autistic, pervasive developmental disorder-not otherwise specified, SLI, or typically developing) from the audio recordings and a suited classification algorithm [8]. The data are from French-speaking participants completing an intonation imitation task, attempting to accurately reproduce perceived lexical and prosodic information. The prompts consist of various types of intonation. Computational tools that have been developed to automatically recognize intonation [9] have also been used to study differential language markers of pathology within this database [10]. The challenge presents an opportunity to discern how separable these pathologies may be along various dimensions such as intonation and rate imitation (which involve both perception and production), rhythm, voice quality, and pronunciation.

The first step in behavioral signal processing is to acquire behavioral signals, while attempting to remit noise sources and maintain ecological validity. It is particularly important to consider and account for other factors of variability in audio, beyond the condition of interest, such as due to recording environment and channel conditions. Recording variations are common in the study of developmental disorders due to austere constraints placed by ecological validity and the difficulties of working with children; the data for the present investigation come from different environments: elementary/high school classrooms (typically developing children) and two different clinics (language impaired populations) [9]. Hence, it is important to account for known sources of variability when conducting further steps in BSP, such as Ringeval et al. (2011) have done by concentrating their study to intonation contours [9].

In this work, we propose a two-fold examination to inform behavioral studies of language impaired children. The first goal is to determine at what level the groups are divisible by this intonation imitation task while considering only features, per the posed challenge task, which are expected to be robust to the particular acoustic variabilities between locations. In particular, we propose the study of prosodic contours and pronunciation quality. Second, we aim to achieve improved classification accuracy using text-independent spectral features which should capture both channel characteristics and voice quality cues.

2. Methodology and Approach

2.1. ASR

We trained an ASR system for phone recognition using Sphinx-3 [11] on the challenge training data set. Speech was parameterized by 13-dimensional MFCCs (plus Δ and $\Delta\Delta$). We use context dependent models with a maximum 32 Gaussians per state in the triphone models. A trigram back-off language model (LM) was trained over French phones using SRILM [12]. The phone error rate (PER) of the development set after Viterbi decoding was 36%. We then perform forced alignment of transcriptions, which are needed for further modeling.

2.2. Prosodic and Formant Templates

The challenge consists of data from children performing a sentence imitation task in which the intonation and modality differ. The children hear a recorded sentence, and reproduce it as closely to what they perceive as they can– therefore, this is both a receptive and expressive task. There are four types of intonation (descending, falling, floating, rising) and four types of modalities (declarative, exclamatory, interogative, imperative). The utterance was replayed if necessary. The recorded data were post-processed to remove false-starts, repetitions, noises from the environment, or speech not related to the task [9].

Both expressive and receptive prosodic impairments have been indicated for children with ASD [13]; thus there is potential to capture atypical prosodic imitation for children with ASD in this task. Accurate recreation of the target statement requires precise use of grammatical and pragmatic prosodic cues. Researchers hypothesize that prosodic deficits attributed to ASD are predominantly pragmatic (e.g., intonation) and affective [14]. Additionally, a previous study on this data suggested that children with ASD were less able to produce the 'rising' contours [10]. It is unclear whether such perceptual and production difficulties exist for children with SLI, but this data offers an opportunity to investigate potential abnormalities.

In order to quantify accurate prosodic imitation, we compute prosodic templates for pitch, intensity, and duration across phones (using forced-alignments), then compare feature contours of each utterance to those templates. Prosodic contour models have been successfully employed for first language learning [15] and ASD research [16]. We also include formant contour templates, since formants have been proposed for modeling dysarthric speaker intelligibility [17]. We assume templates generated using the typically-developing (TD) speakers' recordings in the training data represent optimal reproduction.

To illustrate and further motivate our approach, we generate zero-crossing pitch templates for all four speaker groups for a particular sentence (listed in the caption). The intonation-type in Figure 1 is descending, where the stressed syllable *son* should be more emphasized than the stressed syllable *pa*. We note an ascent for the 'Autism' group in the pitch of the stressed syllables and that *son* is not stressed; the 'PDD-NOS' group exhibits equal emphasis; and the 'SLI' and 'TD' groups exhibit descending emphasis. Here, the 'Autism' intonation is least like 'TD'.

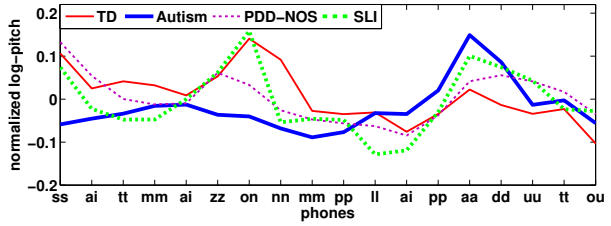

Figure 1: *Normalized log-pitch zero-crossing contours of four groups for the sentence 'Cette maison ne me plait pas du tout.'.*

Contours are constructed across phones (each consecutive phone represents a point in time) with features computed within the boundaries of the corresponding phones. Contours for log-pitch, formants (F1-F3), and intensity are composed of second-order polymial coefficients (curvature, slope, and zero-crossing) fit to each phone; therefore we have three contours for each listed feature. The duration contour is simply the duration of each phone. In our experiments, templates are computed per-sentence as the median feature value for each phone across all considered utterances; only utterances from the TD speakers in the train set are used. We compute two features between template and contour: (1) correlation, as suggested by Duong et al. [15]; and (2) mean absolute difference, or the L1-norm.

2.3. Pronunciation Quality

Articulatory difficulties may exist more so in certain language-impaired groups than others, and thus we investigate pronunciation quality. The hypothesis of apraxia of speech in ASD has been disputed [18], and the prevalence of articulatory difficulties in SLI is not well-researched. However, we have initially observed a poorer pronunciation quality for children in the language-impaired groups (see Section 2.4), and thus expect some divisibility to be observed through classification. We consider variations of the goodness of pronunciation algorithm [19], which is a standard for interactive education systems because it offers a veritable measure of speech fluency.

The goodness of pronunciation (GOP) score [19] computes the following average log-posterior probability of each reference phone p from the output of an ASR system:

$$GOP(p) = |\log P(p|\mathbf{o}^p)|/NF(p) \quad (1)$$

where $\mathbf{o}^p$ is the acoustic observation sequence for phone p and $NF(p)$ are the corresponding number of frames. We compute the exact posterior probability $P(p|\mathbf{o}^p)$ using the forward-backward algorithm. We generated phone lattices using our ASR system and then used *lattice-tool* from SRILM [12] to do the forward-backward computation. *lattice-tool* generates a confusion network consisting of a sequence of sets of competing phones with their posterior probabilities, also known as confusion bins. We then computed the GOP score using the reference phone sequence; the GOP score sequence is used to perform classification per sentence.

2.4. Spectral Energy and Smoothness Features

Some prior literature on speech style and quality in ASD [20, 21, 22] and our own observations implicate spectral cues to offer important distinguishing cues across the various sub-populations of interest. TD children seemed to utter sentences more clearly, while some subjects with language-impairments seemed to mumble and pronounce words with smaller intonational variation (emphasis). In particular, we found that recordings of typically developing children tend to have significantly higher mean utterance intensity values than those of the language-impaired children ($p < 0.01$). Furthermore, we observed that some children with language-impairments showed irregularities in speech fluency resulting in noticeable discontinuities in the spectrogram. However, it should be noted that spectral characteristics may be also influenced by factors other than the health conditions of interest of this challenge, such as environmental and recording conditions.

To explore this further we selected a candidate set of spectral features for a sub-study. Specifically, we chose to use multiresolution, long term spectral features that reflect dynamic variations inspired by our previous studies on speech activity detection. We extracted 360 features that capture spectrogram energy levels and variations. Our features contain total signal

energy, mean and relative energy changes over multiple time scales and frequency bands, and the frequencies with the majority of energy content. We used different energy representations such as intensity and voicing probability. First-order derivatives were calculated in order to account for the temporal dynamics of the utterance and capture possible non-smooth transitions of the spectrogram. We computed long-term functionals of these features including: mean, median, upper and lower quantiles, and differences from the quantile values, producing of functional-value time-series. We included MFCC and RASTA-PLP [23] features for a total of 386 features.

2.5. Experimental Setup

We performed classification using prosodic-template and goodness of pronunciation features with a support vector machine (SVM) classifier [24]. Since the classification metric is unweighted average recall (UAR), the cost-function for each class was weighted by the inverse posterior of the class in order to optimize directly for UAR. Since these methods require the utterance to be known, utterance recognition is developed on the development set as an potential preliminary step to pathology classification, achieving 85% accuracy (chance = 4%).

Our experimental setup for frame-level spectral energy features consisted of a forward feature selection (FFS) approach and a k-nearest neighbor (k-NN) classifier. Class priors were tuned for UAR, and feature selection was optimized on the development set. The per-frame class posteriors were cubed, then averaged over all frames to obtain the utterance-level decision. The cubical probability transformation was empirically found to give better results, suggesting that more salient frames should be given more weight in the final decision.

3. Results

Results on the development set are presented in Table 1 for the 2-class and 4-class tasks. The development set baseline results, which are achieved using 6,373 global functionals from openSMILE [25], SVM, and synthetic sampling to balance classes, are well above chance, especially for the 2-class task (92.8%).

The global prosodic functionals are those features computed without phone-boundary information. Per-sentence classification using total utterance duration achieves 2-class unweighted-average-recall (UAR) of 61.4% and 4-class UAR of 29.6%. Performance decreases with the number of phones in a sentence ($r_{s,4way}$=-0.51, p<0.01).

The template-based models incorporate knowledge of the phone-boundaries from forced-alignment. Although features are generated per-sentence, we pool features from all sentences for classification. Performance with pitch-based templates is significantly above chance. We obtain higher UAR with duration templates than pitch templates or total duration classifiers. Contrary to the trend reported for total duration, UAR for duration templates increases with the number of phones in a sentence ($r_{s,4way}$=0.42, p<0.05). This suggests that having longer utterances provides more information from which to perform classification. Fused pitch and duration template-based features obtained 70% 2-class UAR and 40% 4-class UAR, well above chance, but also well below the baseline results.

Energy-based template models are examined, although potential effects due to channel conditions are unknown. Formant templates, which may relate to intelligibility, show performance similar to that for pitch templates. Although intensity templates obtain UAR only beaten by the duration template models, fusion with the other templates models indicates that intensity templates contain complementary information.

Our goodness of pronunciation features performed above chance in both tasks, suggesting we are capturing disparities in pronunciation quality between groups. However, differences in recording conditions may be captured in the acoustic models.

Table 1: *Unweighted average recalls (%) with the proposed 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

Results for the spectral energy features case study indicate they indeed offer useful discrimination for the classification problem as posed in the challenge, matching or exceeding the baseline result. It should be noted that these features mostly resemble the baseline features, and that feature selection was performed to optimize performance on the development set. However, this performance is reached with only five features for the 2-class task and seven features for the 4-class task, suggesting the features are very informative. Nevertheless, it is unclear still if, and how much, these spectral variations actually are due to the differences in the health conditions (of interest) as against other influencing factors, and as such these results should be viewed with care. This raises a methodological red flag, and we revisit this potential confound in further detail in Section 4.

Setting aside our investigation of template and pronunciation features, we competed in the challenge. We combined five subsystems: (i) two based on linear-kernel SVMs with baseline features; (ii) two using deep neural networks with baseline features; and (iii) one based on our spectral energy features with k-NN classification. We utilized SMOTE [26] upsampling and a hiearchical classification structure: (i) Typ. vs. Atyp.; (ii) ASD vs. SLI; and (iii) PDD-NOS vs. Autism. Late-fusion led to 60.1% UAR on the test-set; further performing unsupervised speaker-clustering as in [27] pushed accuracy to 60.2% UAR.

4. Variability in Acoustic Environments: Effect on Signal Features

Empirical evidence of environmental and recording conditions differences can indicate appropriate algorithmic choices and assist in interpreting results. We observed, through informal listening, a common, distinct reverberation in the typically developing (TD) data compared to the language impaired (LI) data recordings. Blind reverberation estimation is a non-trivial task, and we were unsuccessful at directly quantifying this aspect of room acoustics. Instead, we find differences in the long-term average spectrum (LTAS) of the recordings.

The mean normalized LTAS of all sub-populations are plotted over the frequencies 0-1600 Hz in Figure 2. Differences between groups appear below 600 Hz, more so below 400 Hz. All spectra have spikes of varying height near 100 Hz, possibly an electric hum harmonic. The energy in the LI groups' audio recording spectra below 400 Hz is higher on average, and more diverse than the mean TD LTAS. Furthermore, the typical population has noticeable, smoothed peaks around 230 Hz and 460 Hz, which do no appear in the other spectra. The differences are further evidenced through a classification task. A single gaussian was trained on the LTAS of audio recordings from each group, using only the training set. Then, maximum-likelihood decisions were made for each utterance in the development set.

Firstly, targeting the normalized energy bins of 0-400 Hz, 79.7% 2-way (below baseline) and 51.4% 4-way (ties base-

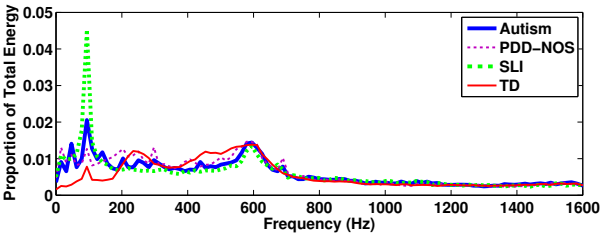

Figure 2: *Mean normalized LTAS of recordings of each group.*

line) unweighted average recall (UAR) were obtained. 'Autism' and 'PDD-NOS' LTAS were very similar, and this may have degraded UAR. In particular, binary classification between 'Autism' and 'PDD-NOS' groups was below chance, 44%, whereas classification between 'SLI' and 'Autism' or 'SLI' and 'PDD-NOS' was above chance at 78% and 68%, respectively. We also performed 4-way classification on the LTAS regions below 400Hz using the audio determined to be silence by the forced-alignment, reaching 41% UAR; however, many factors relating to acoustic modeling and data post-processing may contribute to this performance. Secondly, classification using the bins of 400-8000 Hz was 34% 4-way UAR, below performance with the low-frequency energy.

Since all groups spoke the same utterances, long-term spectral characteristics could reflect room acoustics and voice quality characteristics, as opposed to lexical content. Given that audio file length is short, ranging from 170 ms to 7.2 s (mean = 1.4 s), accurate voice quality features should be difficult to obtain reliably with a global method like LTAS. Additionally considering the classification performance using low-frequency energies and detected silence regions, we argue that the LTAS are mainly reflective of room acoustics.

Another indication of potential channel effects comes from the baseline classification system. Six of the 10 features from the training and development sets most correlated with typical vs. atypical labels overlap– this is astounding given the large, inter-correlated feature set. The six features comprise four flatness functionals on spectral energy-based features (energy, delta-energy, loudness, spectral flux) and two first quartiles of spectral energy-based features (loudness, spectral flux). It is clear that these features will be affected by differences in recording conditions. Although the precise cause and scope of channel effects is unknown, we suggest that variations in recording environments do exist and will influence the results, and hence adjust interpretation of our study in accordance.

5. Discussion of Results

One of the most important questions to consider is, "What kind of performance is expected based on prior literature and experimentation?". Prosodic difficulties are prevalent in children with autism spectrum disorders (ASD), but there has been no such finding in children with specific language impairment (SLI). There is also much debate as to whether children with ASD have articulatory issues. While there is some support for voice quality differences in ASD [20, 21], accurate measurement of atypical voice quality from a single, short utterance is challenging– especially when considering the potential population prevalence of atypical voice quality. Therefore, it is reasonable to expect overall classification accuracy to be well below perfect.

The spectral-energy and smoothness features match the development set baseline in 2-class accuracy (UAR), but exceed it in 4-class accuracy. Given the high performance of these spectral energy features that likely have prominent channel effects due to multiple recording locations (Section 4), we suspect that both our spectral-energy and smoothness features and the baseline features are corrupted by channel effects.

Under the hypothesis of channel effects, we should consider which features are most reliable. However, it is important to note that we are uncertain of the extent of recording artifacts. For example, if there are non-linear channel effects, the signals may be affected in both time and frequency.

Most promising, we find that pitch and duration template models combine for accuracies well above chance. Pitch-tracking and forced-alignment can be affected by channel conditions, but we propose that these features are the most robust to channel effects. While the performance is well below perfect classification (Table 1), it is uncertain how close these numbers are to optimal separability in this task design.

We expected the energy-based template models (i.e., formants and intensity) to potentially have significant channel effects. Formant tracking may have been effected by different acoustic conditions, while intensity is an energy-based feature with many potential confounds described in Section 4. The performance of formant contours suggest there is not major channel effects, indicating we may be capturing some intelligibility factor between groups. In the particular case of intensity contours we found the L1 distance to be discriminative while the correlation feature was not, suggesting that variance in intensity contours was most informative. Therefore, we are uncertain if the variance was due to group differences or feature corruption.

Our implementation of goodness of pronunciation achieved above chance accuracies, indicating some disparities in pronunciation quality are captured between groups. We are uncertain of the potential effects in the acoustic models, especially considering some of the data had a reverberant quality.

We also proposed that by investigating features that may be robust to channel effects, this study may inform future research in speech of children with pervasive developmental disorders. We observed that pitch and duration template models achieved UAR well above chance, indicating that some differences exist between the sub-populations in this imitation task. The confusion matrix for these features tentatively indicates that TD and SLI were most differentiable from other groups, however such conclusions from classification confusion matrices are prone to many influencing factors. We do not see the trends to indicate the Autism group is generally as different from the TD group as shown in Figure 1. Therefore, the performance differences between populations are unclear from our study.

6. Conclusion and Future Work

We presented classification results on the development set using a variety of methods motivated by the task design. We achieved above chance accuracies by using prosodic template modeling and pronunciation quality modeling. The highest accuracy was obtained using spectral amplitude features, matching or exceeding the development set baseline. Investigation of the channel recording conditions suggested differences in recordings between groups recorded at different locations. Coupled with the surprisingly high classification accuracy of the spectral-energy methods, we propose that energy-based methods are not (directly) suitable for this challenge data set.

We hence propose study of features within these domains that may inform realistic separability between groups as well as have the potential to be used for behavioral intervention and monitoring. Further study will investigate which sentences provide the highest distinction between groups.

7. Acknowledgements

This research was supported by funds from NSF and NIH.

8. References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1–31, 2013.

[2] *Diagnostic and Statistical Manual of Mental Disorder, Ed. 4 text revision*, American Psychiatric Assoc., Washington D.C., 2000.

[3] C. A. Baltaxe and J. Simmons, "Prosodic development in normal and autistic children," *Communication problems in autism*, pp. 95–125, 1985.

[4] L. B. Leonard, *Children with specific language impairment*. MIT press, 2000.

[5] G. Conti-Ramsden, Z. Simkin, and N. Botting, "The prevalence of autistic spectrum disorders in adolescents with a history of specific language impairment (sli)," *Journal of Child Psychology and Psychiatry*, vol. 47, no. 6, pp. 621–628, 2006.

[6] J. McCann and S. Peppe, "Prosody in Autism Spectrum Disorders: A Critical Review," *Int. J. Lang. Comm. Dis.*, vol. 38, pp. 325–350, 2003.

[7] B. Wells and S. Peppe, "Intonation abilities of children with speech and language impairments," *Journal of Speech, Language and Hearing Research*, vol. 46, no. 1, p. 5, 2003.

[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistic challenge: Social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[9] F. Ringeval, J. Demouy, G. Szaszak, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1328–1342, 2011.

[10] J. Demouy, M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Périsse, D. Chauvin, S. Viaux, B. Golse, D. Cohen *et al.*, "Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment," *Research in Autism Spectrum Disorders*, vol. 5, no. 4, pp. 1402–1412, 2011.

[11] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 1, pp. 35–45, 1990.

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[13] S. Peppe, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and Expressive Prosodic Ability in Children with High-Functioning Autism," *J. of Speech & Hearing Research*, vol. 50, pp. 1015–1028, 2007.

[14] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and Prosody Characteristics of Adolescents and Adults with High-Functioning Autism and Asperger Syndrome," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 1097–1115, 2001.

[15] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, p. 14, 2011.

[16] J. P. van Santen, E. T. Prud'hommeaux, L. M. Black, and M. Mitchell, "Computational prosodic markers for autism," *Autism*, vol. 14, no. 3, pp. 215–236, 2010.

[17] J. van Santen, X. Niu, J.-P. Hosom, and A. Kain, "Towards automated measures of speech intelligibility in dysarthria," ASHA presentation, 2007, downloaded from http://www.cslu.ogi.edu/people/hosom/pubs/vanSanten-ASHA07-intellDys_2007.pdf, on 3-19-2013.

[18] L. D. Shriberg, R. Paul, L. M. Black, and J. P. van Santen, "The hypothesis of apraxia of speech in children with autism spectrum disorder," *Journal of autism and developmental disorders*, vol. 41, no. 4, pp. 405–426, 2011.

[19] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[20] S. J. Sheinkopf, P. Mundy, D. K. Oller, and M. Steffens, "Vocal atypicalities of preverbal autistic children," *Journal of Autism and Developmental Disorders*, vol. 30, no. 4, pp. 345–354, 2000.

[21] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proc. Interspeech*, 2012.

[22] A. McAllister, J. Sundberg, and S. R. Hibi, "Acoustic Measurements and Perceptual Evaluation of Hoarseness in Children's Voices," *Logopedics Phoniatrics Vocology*, vol. 23, 1998.

[23] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 121–124.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622407.1622416

[27] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors," in *Proceedings of InterSpeech*, Sep. 2012.