Robust and Accurate Features for Detecting and Diagnosing Autism Spectrum Disorders

Meysam Asgari, Alireza Bayestehtashk and Izhak Shafran

Center for Spoken Language Understanding
Oregon Health & Science University
Portland, OR, USA
{asgari, shafrani, bayesteh}@ohsu.edu

10.21437/Interspeech.2013-65

Abstract

In this paper, we report experiments on the Interspeech 2013 Autism Challenge, which comprises of two subtasks – detecting children with ASD and classifying them into four subtypes. We apply our recently developed algorithm to extract speech features that overcomes certain weaknesses of other currently available algorithms [1, 2]. From the input speech signal, we estimate the parameters of a harmonic model of the voiced speech for each frame including the fundamental frequency (f_0). From the fundamental frequencies and the reconstructed noise-free signal, we compute other derived features such as Harmonic-to-Noise Ratio (HNR), shimmer, and jitter. In previous work, we found that these features detect voiced segments and speech more accurately than other algorithms and that they are useful in rating the severity of a subject's Parkinson's disease [3]. Here, we employ these features, along with standard features such as energy, cepstral, and spectral features. With these features, we detect ASD using a regression and identify the sub-type using a classifier. We find that our features improve the performance, measured in terms of unweighted average recall (UAR), of detecting autism spectrum disorder by 2.3% and classifying the disorder into four categories by 2.8% over the baseline results.
Index Terms: speech analysis, autism spectrum disorder

1. Introduction

Autism spectrum disorder (ASD) cover a range of developmental disabilities that can cause significant social, communication, and behavioral challenges. Children with ASD often are self-absorbed in their private world and they have difficulty communicating and interacting with others. While not every child with ASD has a language problem, the majority have difficulty using language effectively, especially when conversing with others. Often they exhibit unusual pitch and intonation, for example, monotonous pitch, reduced stress, odd rhythm, large pitch range [4], and even differences in harmonic structure of their speech [5]. There has been continual interest in characterizing these variations in ASD and potentially exploit them in objectively quantify and categorizing the language impairments in ASD.

The range of disorders in ASD can be categorized according to Diagnostic and Statistical Manual of Mental Disorders (DSM), published by American Psychiatric Association. Most clinicians in the US follow the fourth edition (DSM-IV) [6]. The diagnostic category pervasive developmental disorders (PDD) refers to disorders characterized by delays in the development of multiple basic functions including socialization and communication. This category includes Asperger and Rett syndromes. Pervasive developmental disorder not otherwise specified (PDD-NOS) is one of the five ASDs, characterized as "severe and pervasive impairment in the development of reciprocal social interaction or verbal and nonverbal communication skills, or when stereotyped behavior, interests, and activities are present, but the criteria are not met for a specific PDD" or for several other disorders. Unrelated to the above conditions, a child could suffer from limited ability to socialize and communicate, not because of general developmental disorders, but due to specific language impairments such as dysphagia. In all these cases, prosody and intonation are compromised perhaps in different ways, and that is a topic of considerable research interest currently especially for developing useful intervention strategies.

In this paper, we report our experiments on the *Autism Sub-Challenge* of Interspeech 2013. The challenge consists of two tasks: 1) a binary 'Typicality' classification task with classes – TYPicality developing (TYP) and ATYpically developing (ATY), and a four-way 'Diagnosis' task for classifying children into 4 categories – TYP, PDD, PDD-NOS, and specific language impairment such as DYSphasia (DYS). The paper is organized as follows. We start the harmonic model of voiced speech and our feature extraction algorithms in Section 2. Our experiments and the results are reported in Section 4. Finally, we conclude with summary of our key results.

2. Speech Analysis Using Harmonic Model

The popular source-channel model of voiced speech considers glottal pulses as a source of period waveforms which is being modified by the shape of the mouth assumed to be a linear channel. Thus, the resulting speech is rich in harmonics of the glottal pulse period.

2.1. Harmonic Model

The harmonic model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related, that is, the frequencies of the sinusoids are multiples of the fundamental frequency. This model is tailored to capture the rich harmonic nature of voiced segments in speech.

Stylianou introduced a Harmonic plus Noise Model (HNM) for speech analysis and synthesis, in which speech signals are represented as a time-varying harmonic component plus a modulated noise component [7]. The harmonic part accounts for the periodic component of the speech signal while the noise part accounts for its non-periodic components. Speech decomposition using a HNM is useful for applications in speech synthesis,

Copyright © 2013 ISCA

voice conversion, speech enhancement, and speech coding.

Let $\mathbf{y} = [\mathbf{y(t_1)}, \mathbf{y(t_2)}, \ldots, \mathbf{y(t_N)}]^{\mathbf{T}}$ denote the speech samples in a voiced frame, measured at times $t_1, t_2, \ldots, t_T$. The samples can be represented with a harmonic model with an additive noise $\mathbf{n} = [\mathbf{n(t_1)}, \mathbf{n(t_2)}, \ldots, \mathbf{n(t_N)}]^{\mathbf{T}}$ as follow:

$$\begin{aligned} s(t) &= a_0 + \sum_{h=1}^{H} a_h cos(2\pi f_0 ht) + b_h sin(2\pi f_0 ht) \\ y(t) &= s(t) + n(t) \end{aligned} \quad (1)$$

where H denotes the number of harmonics and $2\pi f_0$ stands for the fundamental angular frequency. The harmonic signal can be factored into coefficients of basis functions, α, β, and the harmonic components which are determined solely by the given angular frequency $2\pi f_0$ and the choice of the basis function $\psi(t)$.

$$\begin{aligned} s(t) &= \begin{bmatrix} 1 & A_c(t) & A_s(t) \end{bmatrix} \begin{bmatrix} a_0 \\ \alpha \\ \beta \end{bmatrix} \\ A_c(t) &= \begin{bmatrix} \cos(2\pi f_0 t) & \cdots & \cos(2\pi f_0 Ht) \end{bmatrix} \\ A_s(t) &= \begin{bmatrix} \sin(2\pi f_0 t) & \cdots & \sin(2\pi f_0 Ht) \end{bmatrix} \end{aligned} \quad (2)$$

Stacking rows of $[1 \;\; A_c(t) \;\; A_s(t)]$ at $t = 1, \cdots, T$ into a matrix $\mathbf{A}$, equation (2) can compactly represented in matrix notation as:

$$\mathbf{y} = \mathbf{A}\,\mathbf{m} + \mathbf{n} \quad (3)$$

where $\mathbf{y} = \mathbf{A}\,\mathbf{m}$ corresponds to a expansion of the harmonic part of voiced frame in terms of windowed sinusoidal components, and $\Theta = [f_0, \mathbf{b}, \sigma_n^2, H]$ is the set of unknown parameters.

2.2. Parameter Estimation

Assuming the noise samples **n** are independent and identically distributed random variables with zero-mean Gaussian distribution, the likelihood function of the observed vector, y, given the model parameters can be formulated as following equation. The parameters of vector **m** can then be estimated by maximum likelihood (ML) approach.

$$\begin{aligned} \mathbf{L}(\Theta) &= -\frac{D}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}||\mathbf{y} - \mathbf{Ab}||^2 \\ \hat{\mathbf{m}}_{ML} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y} \end{aligned} \quad (4)$$

Under the harmonic model, the reconstructed signal $\hat{\mathbf{s}}$ is given by $\hat{\mathbf{s}} = \mathbf{A}\,\mathbf{m}$. So far, we assumed that the pitch f_0 was given. However, in practice, the pitch needs to be estimated. It can be computed by maximizing the energy of the reconstructed signal over the pre-determined grid of discrete f_0 values ranging from $f_{0\,min}$ to $f_{0\,max}$.

$$\hat{f}_{0\,ML} = \arg\max_{f_0} \; \hat{\mathbf{s}}^T\hat{\mathbf{s}} \quad (5)$$

2.3. Segmental Pitch Tracking

The pitch variations are inherently limited by the motion of the articulators in the mouth during speech production and hence they cannot vary arbitrarily between adjacent frames. This smoothness constraint can be enforced using a first order Markov dependency between pitch estimates of successive frames. Adopting the popular hidden Markov model framework, the estimation of pitch over utterances can be formulated as follows. Let $\mathbf{Y} = \{\mathbf{y}_0, \ldots, \mathbf{y}_M\}$, and $\mathbf{F}_0 = \{f_0^{(0)}, \ldots, f_0^{(M)}\}$ be M length sequences of observed frames and candidate pitch estimates respectively.

$$\hat{\mathbf{F}}_0 = \underset{\mathbf{F}_0}{\text{argmax}}\, P(\mathbf{F}_0|\mathbf{Y}) = \underset{\mathbf{F}_0}{\text{argmax}}\, P(\mathbf{Y}|\mathbf{F}_0)P(\mathbf{F}_0)$$

The observation probabilities are assumed to be independent given the hidden states or candidate pitch frequencies here. A zero-mean Gaussian distribution defined over the pitch difference between two successive frame is a reasonable approximation for the first order Markov transition probabilities [8], $P(f_0^{(i)}|f_0^{(i-1)}) = \mathcal{N}(f_0^{(i)} - f_0^{(i-1)}, \sigma_t^2)$. Putting all this together and substituting the likelihood from the Equation 5, the pitch over an utterance can be estimated as follows.

$$\hat{\mathbf{F}}_0 = \underset{\mathbf{F}_0}{\text{argmax}} \left[\sum_{i=0}^{M} \hat{\mathbf{s}_i}^T \hat{\mathbf{s}_i} | f_0^{(i)} + \log \mathcal{N}(f_0^{(i)} - f_0^{(i-1)}, \sigma_t^2) \right] \quad (6)$$

Thus, the estimation of pitch over an utterance can be cast as an HMM decoding problem and can be efficiently solved using Viterbi algorithm.

2.4. Jitter and shimmer

Jitter and shimmer refer to a short-term (cycle-to-cycle) perturbation in the f_0 and the amplitude of voice waveform respectively. Perturbation analysis is based on the fact that small fluctuations in frequency, and amplitude of waveform reflect the inherent noise of voice. These measures can be sensitive to noise. We alleviate this problem by estimating jitter and shimmer from the signal reconstructed using the estimated parameters of the harmonic model [3].

2.4.1. Shimmer

In order to compute shimmer, we first represent the speech waveform using the harmonic model with time-varying amplitudes (HM-VA) as shown in equation 7 [9].

$$\begin{aligned} s(t) &= a_0(t) + \sum_{h=1}^{H} [a_h(t)\cos(2\pi f_0 ht)] \\ &+ \sum_{h=1}^{H} [b_h(t)\sin(2\pi f_0 ht)] \end{aligned} \quad (7)$$

Note, this is different from the harmonic model represented previously in Equation 1. Unlike, the previous model whose harmonic coefficients are fixed, in the time-varying model, as the name implies, the coefficients are allowed to vary $a_h(t)$ and $b_h(t)$ over time. Thus, this model is capable of capturing sample to sample variation in harmonic amplitude within a frame. Given the limitations of the articulators, it is reasonable to assume that the sample to sample variation is smooth. This can be represented as a superposition of small number of basis functions ψ_i as in equation 8 [9].

$$a_h(t) = \sum_{i=1}^{I} \alpha_{i,h}\psi_i(t) \;\;, \;\; b_h(t) = \sum_{i=1}^{I} \beta_{i,h}\psi_i(t) \quad (8)$$

We represent this smoothness constraints within a frame using four $(I = 4)$ Hanning windows as basis functions. For a frame of length M, the windows are centered at 0, $M/3$, $2M/3$, and M. Each basis function is $2M/3$ samples long and has an overlap of $M/3$ with immediate adjacent window. The parameters of this model can be expressed, once again, as a linear model,

similar to Equation 3, but this time the A and m have dimensions four times the original dimensions. Given the fundamental frequency from 6, we compute $a_h(t)$ and $b_h(t)$ using a maximum likelihood framework.

Shimmer can be considered as a function $f(t)$ that scales the amplitudes of all the harmonics in the time-varying model.

$$c_h(t) = c_h f(t) + e(t), \quad t = 1, \ldots, T\ , h = 1, \ldots, H \quad (9)$$

where $c_h = \sqrt{\sum_{h=1}^{H} a_h^2 + b_h^2}$ denotes the amplitude of the harmonic components in harmonic model with constant amplitudes and $c_h(t)$ is the counterpart from the time-varying model. Once again, assuming uncorrelated noise, f(t) can be estimated using maximum likelihood criterion.

$$\hat{f}(t) = \frac{\sum_{h=1}^{H} c_h c_h(t)}{\sum_{h=1}^{H} c_h^2} \quad (10)$$

The larger the tremor in voice, the larger the variation in $f(t)$. Hence, we use the standard deviation of $f(t)$ as a summary statistics to quantify the shimmer.

2.4.2. Jitter

Given an estimate pitch period of the frame, we first create a matched filter by excising a one pitch period long segment from the signal estimated with the harmonic model from the center of the frame. This matched filter is then convolved with the estimated signal and the distance between the maxima defines the pitch periods in the frame. The perturbation in period is normalized with respect to the given pitch period and its standard deviation is an estimate of jitter.

2.5. Harmonic-to-noise ratio (HNR)

Researchers have used HNR in the acoustic studies for the evaluation of voice disorders. Given the reconstructed signal as the harmonic source of vocal tract, the noisy part is obtained by subtracting the reconstructed signal from the original speech signal. The noisy part encompasses everything in the signal that is not described by harmonic components including the frication noise, the waveform fluctuations, etc. HNR and the ratio of energy in first and second harmonics (H12) can be computed from the HM-VA as follow.

$$\begin{aligned} c_h(t) &= \sqrt{\sum_{i=1}^{I} a_h(t)^2 + b_h(t)^2} \\ HNR &= \log \sum_{t=1}^{N} \sum_{h=1}^{H} c_h(t)^2 - \log \sum_{t=1}^{N} \left(y(t) - s(t)\right)^2 \\ H12 &= \log \sum_{t=1}^{N} c_1(t)^2 - \log \sum_{t=1}^{N} c_2(t)^2 \end{aligned} \quad (11)$$

3. Corpus

Empirical evaluation reported in this paper were performed on "Child Pathological Speech Database" (CPSD) [1] collected from 99 children, age 9 to 18, through two hospitals located in Paris, France. This dataset provides 2542 short speech utterances collected for assessing children's abilities in imitation of different types of prosody contours. Based on the prosodic dependencies of French language, sentences carry out 4 intonations type including *descending, falling, floating, and rising*. Subjects, were asked to read 26 phonetically easy sentences and they were recorded in separate files. As a clinical reference, the severity of subjects condition were measured by clinicians using the DSM-IV criteria [6], where 35 of these children showed PDD either of Autism Spectrum Condition (ASC, 12 children), specific language impairment (SLI, 13 children) or PDD non-otherwise specified (PDD-NOS, 10 children). The corpora includes rich annotation such as speaker meta-data, orthographic transcript, phonemic transcript, and segmentation. Also, the corpus treats sentences read by the same speaker as independent samples partitioned randomly in test, development, and training sets.

Table 1: Unweighted average recall (UAR) for detecting ASD kids from typically developing (TD) kids, and for classifying the ASD kids into four sub-types.

Speech Features	ASD Tasks	
	ASD vs. TD	4-subtypes of ASD
Baseline	90.7	67.1
Improved Features	**93.58**	**69.42**

4. Experiments

4.1. Features

As in most speech processing systems, we extract 25 millisecond long frames using a Hanning window at a rate of 100 frames per second before computing the frame-level features. Voicing related features including pitch, HNR, the ratio of energy in first to second harmonics (H12), jitter, and shimmer are derived from the expressed harmonic analysis over the voiced frames.The features computed at the frame-level needs to be summarized into a global feature vector of fixed dimension for each read sentence. Each feature was summarized across all frames from the voiced segments in terms of standard distribution statistics such as mean, median, variance, minimum and maximum. We also computed the covariance matrix (upper triangular elements) of frame-level feature vectors over voiced segments to capture interaction between features.The resulting per-sentence voice quality feature vector was later augmented by per-sentence energy, spectral, and cepstral related features provided from baseline. For more detail regarding the baseline frame-level features and also functionals that are applied to those feature, we refer the reader to the challenge paper [10].

4.2. Regression and classification models

Typically, in clinical applications, the class distributions are highly unbalanced, as it is in the four subtypes within this corpus. The challenge evaluation metric of unweighted average recall attempts to normalize the influence of the highly skewed classes. We employed a support vector classifier and a support vector regression respectively to detect ASD cases and to identify the subtypes. Both the regression and classifer were learned from the data using open-source WEKA toolkit[11]. For the training the regression and classifier, we retained the hyper parameters from the baseline system, $C = 0.001$. For the test set, all labeled data from train and developing sets were pooled for training and a new model learned using parameters reported in the baseline. Since the class distribution in the training data was skewed, we upsampled instances in atypicality categories

(PDD, NOS, and DYS) by using a factor of five. We refer the reader to the baseline challenge paper for more detail [10]. Table 1 reports UAR evaluated from baseline feature vectors and proposed feature vector on detecting ASD and classifying the sub-types. From the results, it is clear that our voice quality related features (derived by harmonic analysis) significantly improve UAR in both tasks.

5. Conclusions

In summary, we considered several speech measures to detect children with ASD and to classify them into four subtypes. For both tasks, our features can be categorized into four groups – voice quality features (estimated from harmonic analysis), energy-related features, spectral features, and cepstral features. We found that our features, specifically the voice quality features, improve the performance of both tasks in terms of unweighted average recall (UAR).

6. Acknowledgements

This research was supported in part by NIH Award AG033723, NSF Awards 1027834, 0964102 and 0905095 and Google Faculty Award. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH or NSF.

7. References

[1] Fabien Ringeval, Julie Demouy, Gyorgy Szaszak, Mohamed Chetouani, Laurence Robel, Jean Xavier, David Cohen, and Monique 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.

[2] Meysam Asgari, Izhak Shafran, and Alireza Bayestehtashk, "Robust detection of voiced segments in samples of everyday conversations using unsupervised HMMs," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 438–442.

[3] Meysam Asgari and Izhak Shafran, "Extracting cues from speech for predicting severity of Parkinson's disease," in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 462–467.

[4] Kathleen Hubbard and Doris A Trauner, "Intonation and emotion in autistic spectrum disorders," *Journal of psycholinguistic research*, vol. 36, no. 2, pp. 159–173, 2007.

[5] Yoram S Bonneh, Yoram Levanon, Omrit Dean-Pardo, Lan Lossos, and Yael Adini, "Abnormal speech spectrum and increased pitch variability in young autistic children," *Frontiers in human neuroscience*, vol. 4, 2010.

[6] Samuel B Guze, "Diagnostic and statistical manual of mental disorders, (dsm-iv)," *American Journal of Psychiatry*, vol. 152, no. 8, pp. 1228–1228, 1995.

[7] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," in *Ph.D. dissertation, Ecole Nationale des Tlcomminications*, 1996.

[8] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76 – 87, 2004.

[9] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 2, pp. 1769–72.

[10] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013.

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.