HHS Public Access

Author manuscript

Annu Int Conf IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2021 July 01.

Published in final edited form as:

Annu Int Conf IEEE Eng Med Biol Soc. 2020 July ; 2020: 6111–6114. doi:10.1109/EMBC44109.2020.9175264.

Automatic Assessment of Language Ability in Children with and without Typical Development

Robert Gale[1], **Jill Dolata**[1], **Emily Prud'hommeaux**[2], **Jan van Santen**[1], **Meysam Asgari**[1]

[1]Faculty of Pediatrics Oregon Health & Science University, Portland, OR 97239 USA

[2]Emily Prud'hommeaux is with the Department of Computer Science, Boston College, Boston, MA 02467, USA

Abstract

This study describes a fully automated method of expressive language assessment based on vocal responses of children to a sentence repetition task (SRT), a language test that taps into core language skills. Our proposed method automatically transcribes the vocal responses using a test-specific automatic speech recognition system. From the transcriptions, a regression model predicts the gold standard test scores provided by speech-language pathologists. Our preliminary experimental results on audio recordings of 104 children (43 with typical development and 61 with a neurodevelopmental disorder) verifies the feasibility of the proposed automatic method for predicting gold standard scores on this language test, with averaged mean absolute error of 6.52 (on a observed score range from 0 to 90 with a mean value of 49.56) between observed and predicted ratings.

Clinical relevance—

We describe the use of fully automatic voice-based scoring in language assessment, including the clinical impact this development may have on the field of speech-language pathology. The automated test also creates a technological foundation for the computerization of a broad array of tests for voice-based language assessment.

I. INTRODUCTION

Since untreated language disorder – a disorder with a prevalence of at least 7% [1] – can lead to behavioral and educational problems, early language assessment is needed for identification of language disorders as well as for intervention planning and progress monitoring. Large-scale screening and assessment efforts can burden staff and resources. Computer-based assessment is a possible solution, and there have been advances in computerized tests of receptive language [2]. Comprehensive evaluation of spoken language in children, however, requires assessment of expressive as well as receptive language. Recent advances in the accuracy of automatic speech recognition (ASR) systems have resulted in swift integration into educational and professional settings [3]; however, ASR capabilities are reduced (i.e., word error rates are increased) when there are speech or

asgari@ohsu.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

language differences, such as accents, speech sound substitutions, distortions, or vocal quality differences (e.g., pitch, tone, prosody) [4].

Most research in the area of “non-standard” speech is on adults (e.g., dysarthria [5] and accents in non-native speakers [6]). It is also very important that these new technologies be accessible to children. Some researchers are beginning to study ASR accuracy for typically developing children [6]; however, still little is known about how ASR systems work for children and even less about how they work for children with neurodevelopmental differences. If ASR can be adopted for any kind of language assessment, it is imperative that it be researched with children who have communication disorders.

This paper describes a novel use of automatic scoring in language assessment, including the clinical impact this development may have on the field. It is possible that this type of technological advancement could improve efficiency and improve access in rural areas. Automated assessment tools may also benefit intervention, making it feasible to perform detailed progress tracking, as required by many insurance providers and school districts.

In addition to these procedural advantages, our automatic approach can also improve upon traditional standardized scoring procedures. The results of a standardized test are usually a highly simplified numerical distillation of that assessment. The risk of information loss is compounded by conventional scoring methods. The conventional approach to a language test is to present the subject with a series of stimuli and score the subject’s responses according to the rules of the subtest (e.g., a count of correct responses), then the raw score is converted to a scaled score based on the normative sample of the test. Similarly, the scaled scores of several assessments are typically summed and then converted to a norm-referenced composite score.

Our approach aims to address these limitations by considering all available information after predicting the score to potentially gain several advantages. First, every item’s contribution to the overall score prediction is automatically weighted, emphasizing the aspects of the test which are most informative, and de-emphasizing those which are least informative. Similarly, we can establish how each item contributes to a final prediction. Second, we are able to “de-noise” the data, statistically smoothing outliers from situations as necessary (e.g., responses that were mistranscribed by the automatic speech recognizer).

II. Methods

A. Participants

Participants for this analysis were part of a larger study designed to develop automated administration and scoring for several expressive language measures. This subset of participants included 104 English-speaking children who completed automatic administrations of the Sentence Repetition Task (SRT). Participants were recruited from university electronic health records, community organizations, and social media. Upon enrollment, study staff reviewed medical and developmental histories with caregivers. Based on prior diagnoses, children were classified as typically developing (TD) or atypically developing (AD) placed in a primary clinical category of autism spectrum disorder (ASD),

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

language impairment (LI), or attention deficit hyperactivity disorder (ADHD). The research described in this study was approved by the Oregon Health & Science University (OHSU) institutional review board (IRB). All research actions were governed by IRB policies, including ethical considerations for human subjects. Consent for this study was obtained from at least one legal guardian. Table I presents means and standard deviations (SD) for age, gender, and gold standard SRT scores.

B. Data Collection

1) Task creation: As part of the effort to create alternative stimuli parallel to those found in published language assessments, we developed two pieces of software: (1) a word lookup tool used by the SLP to identify words with comparable frequency [7], [8], [9], age of acquisition [10], and syllabic structure [11] to the words found in the published subtest stimuli; and (2) a sentence evaluation and analysis tool that ensures that alternative sentences designed by the SLP are equivalent to the subtest sentences in various measures of syntactic complexity [12], [13], reading level, and grade level, as well as mean word frequency, mean age of acquisition, and overall emotional valence [14]. Using these tools, we created a novel replication of the clinical Evaluation of Language Fundamentals, fourth edition [15], a test that was initially developed in 1982 and is currently in its fifth edition.

2) Test administration: Children were given the SRT task using software developed for this project. The iPad application features a voice-based user interface: the tablet speaks to the user (in the form of an on-screen virtual examiner), and the user speaks to the tablet. The tablet application emulates conventional administration as much as possible. The child's response is recorded and scored conventionally (i.e., by study staff). Paper scores and iPad scores were reviewed and discrepancies resolved during consensus meetings that included a licensed speech-language pathologist and/or licensed psychologist. Audio responses were later transcribed by research assistants and graduate students.

C. Sentence Repetition Task

In the SRT, an SLP asks a child to repeat 32 sentences of increasing complexity and length. Responses are scored for semantic and syntactic accuracy on a 0-3 scale: 3 indicates a perfect verbatim response, 2 indicates one error, 1 indicates two or three errors, and a score of 0 indicates four or more errors. We note that a word transposition is scored as one error unless it changes the meaning, such as a subject-object reversal, in which case two errors are counted. Repetitions of words or sounds are not counted as errors. As in many other tests, items increase in difficulty, and to minimize frustration a stopping rule is used after five consecutive zero point responses. Scores were computed for each item conventionally and automatically.

1) Limitations of Conventional Scoring: Conventional scoring constructs a raw score by summing item-level scores across all test items and comparing this raw score with normative data to describe the overall performance of the test taker. However, paying equal attention to test items while summing item-level scores ignores the intrinsic complexity of and difference between presented items. Figure 1 illustrates the probability distribution of true item-level scores (vertical axe) of 32 test items (horizontal axe) across TD (top) and AD

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(bottom) children, plotted separately for each test item using violin plots. The length of sentences (i.e., the number of words) in test items incrementally increases from 6 to 19 words. According to both plots, sentence-level scores gradually drop as a function of item length suggesting that test items do not equally contribute toward the final score.

III. Automatic Assessment of Sentence Repetition Task

A. Automatic Speech Recognition

The first stage in automatic processing of the SRT is to obtain a reliable transcription of the participants' responses. Our methods closely follow our previous work with the RS task [16], built around a hybrid of a time-delay neural network (TDNN) and a hidden Markov model/Gaussian mixture model (HMM-GMM) trained using the Kaldi ASR toolkit [17]. The success of a TDNN model depends heavily on a large quantity of training audio, which is not readily available for children's speech. To overcome this challenge, we first created a base model trained on the Librispeech corpus, consisting of 1000 hours of adult volunteers reading books aloud [18]. To adjust for the acoustic differences between adult and children's speech, we applied a transfer learning technique [19], selectively adapting the model with 1.5 hours of children performing the RS task from the Center for Spoken Language Understanding (CSLU) Autism Corpus [16] as well as 1.2 hours from the 2nd graders subset of the CSLU Kid's Corpus [20] performing a sentence reading task. Using five-fold cross-validation, we were also able to include the target SRT audio recordings in our adaptation data set, training five separate models on a rotating four-fifths of the target data set, then obtaining our transcriptions from the remaining held-out fifth. The new SRT recordings amount to about 8.1 hours of audio. Ideally, a given response would consist of only child's speech, but in reality, the recordings often contain extraneous speech of child such as fillers (e.g., "um") and conversational words. It is also very common to observe examiner's speech in the beginning or the end of recordings encouraging the child with a few words (e.g., "very good!").

B. Sentence-Level Measures

While conventional scoring of RST is basically a function of the number of errors, as described in Section II-C, there are some nuances to the definition of an error. In constructing our statistical model, we implement a simpler definition of error, avoiding for the time being any complex disfluency/repair modeling or semantic analysis. For our persentence error vectors, we used the Levenshtein ((minimum) edit distance) algorithm to align the correct answer to the subject's attempt. From the resulting alignment, we can count the number of correct, omitted, added, and substituted words. The effect of our naive approach to error counting is shown in Figure III-A, a response with a number of repairs and disfluencies. According to the scoring rules, there are only 2 errors, an addition and a substitution. In computing edit distance for this example, while we have configured the ASR to filter out a filler like "um," we do not distinguish between repairs and insertions, and we compute 7 total errors. Similarly, any incorrectly transcribed or superfluous words picked up by the ASR (as mentioned in Section III-A) will be recorded as sentence-level errors, further contributing to the noise in our automatic scoring system.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

C. Subject-Level Measures

To train a machine learning model for predicting the overall score, sentence-level measures are summarized as a global feature vector of fixed dimensions. This can be accomplished by concatenating the four sentence-level measures across all 32 prompts. However, since some participants did not attempt each prompt, we needed to fill in missing values to achieve vectors of fixed dimensions across all subjects. Informed by the test's stopping rules, we input the missing measures with the average measures of the participant's last five responses. Thus we generate a 128-dimensional feature vector (four measures per item) representing a detailed assessment of each participant's vocal responses across all 32 test prompts.

D. Feature Normalization

Prior to training a regression model, we scale the range of computed features into a constrained range by adopting a non-linear transformation, *quantile transform*, from the open-source scikit-learn machine learning toolkit [21]. Our quantile transform maps the observed features onto 50 quantiles of a uniform distribution. This serves two useful purposes in our computational framework. First, it allows our linear regression models to better reflect the non-linear nature of the observed data by spreading out those values observed most often. Second, the transform attempts to squeeze outliers into a more reasonable scale and to prevent those responses with an unusually large number of word errors from distorting the model's predictive power. The quantile transform has also been successfully used in suppressing background noise in noisy speech signals [22].

E. Prediction Models

As a baseline, we implement a rule-based approximation of the standard scoring procedure. Using the sentence-level measures described in Section III-B, we sum the number of errors per prompt. We then give each prompt a score of 0-3 points per the conventional scoring rules described in Section II-C. We also employ several regression models from the scikit-learn toolkit [21]. Not all the extracted features are expected to be informative, and in fact many are likely to be noisy and may increase the risk of overfitting. For the compensation, we adopt two forms of regularization, L2-norm in the ridge regression, and hinge loss function in support vector regression. To compare the performance against a chance model, we adopt a simple regression function (*Dummy*) that correspondss to guessing the same score (mean of true scores) for all subjects. Lastly, to explore the effect of ASR errors on score prediction models, we ran all prediction experiments on both the manually transcribed responses and the ASR-generated transcripts.

IV. Results

To demonstrate the performance of our method we adopt a cross-validation (CV) technique in which the train and test sets are separated and rotated over the entire data set. Given the limited number of subjects in our cohort, we use a leave-one-out (LOO) CV scheme that iteratively trains the model on all subjects except a single subject left for the test. At every iteration, the model predicts the score for the test example and measures the performance in term of mean absolute error (MAE) between true and predicted score values. Finally, the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

overall performance is measured by averaging MAE values across all models. Additionally, we report Pearson correlation coefficients (PCC) between observed and predicted scores across subjects. Using 128-dimensional subject-level features extracted from ASR-generated transcription of vocal responses including both TD and AD children, we evaluated the performance of the several regression models on predicting the SRT scores. The results are reported in Table II for the LOO CV. The *Dummy* model that guesses the mean SRT score on this data incurs an MAE of about 18.14. Table II indicates that SRT scores can be predicted well by the two alternative regression models, better than rule-based scoring, with an MAE of about 4.28 and a PCC of about 0.97 on manual transcription and MAE of about 6.52 and a PCC of about 0.92 on ASR transcription using Linear support vector regression (SVR). As expected, results show that regardless of the regression model employed for the prediction, features extracted from the manual transcriptions outperform those extracted from ASR output. The slightly higher MAE using ASR compared to manual transcription suggests that proposed automatic method using an ASR system can be effectively employed for assessing the SRT.

V. Conclusion

This paper describes an experimental paradigm for administration of an expressive spoken language test, methods for automatically extracting useful features from ASR-generated transcription of vocal responses, and models for predicting the quality of language skills as reflected in clinically obtained test scores. With enough vectors to train a statistical regression model, we were able to make reasonably accurate predictions for the SRT scores. In future work, we intend to evaluate this approach in the remaining subtests, then apply the concept to other subscales and language composites.

Acknowledgments

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Number R01DC013996. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E, O'Brien M. Prevalence of specific language impairment in kindergarten children. Journal of speech, language, and hearing research. 1997;40(6):1245–1260.

[2]. Carson K, Gillon G, Boustead T. Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. Asia Pacific Journal of Speech, Language and Hearing. 2011;14(2):85–101.

[3]. Holland VM, Fisher FP. The Path of Speech Technologies in CALL: Tracking the Science In: The Path of Speech Technologies in Computer Assisted Language Learning. Routledge; 2008 p. 15–32.

[4]. Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, et al. Automatic speech recognition and speech variability: A review. Speech communication. 2007;49(10-11):763–786.

[5]. Young V, Mihailidis A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. Assistive Technology. 2010;22(2):99–112. [PubMed: 20698428]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

[6]. Van Doremalen J, Cucchiarini C, Strik H. Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing. 2009;2010:1–13.

[7]. Davies M The corpus of contemporary American English. BYE, Brigham Young University; 2008.

[8]. Masterson J, Stuart M, Dixon M, Lovejoy S. Children's printed word database: Continuities and changes over time in children's early reading vocabulary. British Journal of Psychology. 2010;101(2):221–242. [PubMed: 20021708]

[9]. Brysbaert M, New B. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior research methods. 2009;41(4):977–990. [PubMed: 19897807]

[10]. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. Age-of-acquisition ratings for 30,000 English words. Behavior research methods. 2012;44(4):978–990. [PubMed: 22581493]

[11]. Weide R The Carnegie Mellon pronouncing dictionary [cmudict. 0.6]. Pittsburgh, PA: Carnegie Mellon University 2005;.

[12]. Dowty DR, Karttunen L, Zwicky AM. Natural language parsing: Psychological, computational, and theoretical perspectives. Cambridge University Press; 2005.

[13]. Yngve VH. A model and an hypothesis for language structure. Proceedings of the American philosophical society. 1960;104(5):444–466.

[14]. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Computational Intelligence. 2013;29(3):436–465.

[15]. Semel E, Wiig E, Secord W. Clinical evaluation of language fundamentals, (CELF-4) The Psychological Corporation San Antonio, TX 2003;.

[16]. Gale R, Chen L, Dolata J, van Santen J, Asgari M. Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques. In: Proc. Interspeech 2019; 2019 p. 11–15.

[17]. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, et al. The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding IEEE Signal Processing Society; 2011 IEEE Catalog No.: CFP11SRW-USB.

[18]. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE; 2015 p. 5206–5210.

[19]. Ghahremani P, Manohar V, Hadian H, Povey D, Khudanpur S. Investigation of transfer learning for ASR using LF-MMI trained neural networks. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2017 p. 279–286.

[20]. Shobaki K, Hosom JP, Cole RA. CSLU: Kids Speech Version 1.1 LDC2007S18. Linguistic Data Consortium; 2007.

[21]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research. 2011;12:2825–2830.

[22]. Farhadloo M, Sayadian A, Asgari M, Mostafavi A. Causal Multi Quantile Noise Spectrum Estimation for Speech Enhancement. In: 2008 Australasian Telecommunication Networks and Applications Conference IEEE; 2008 p. 112–115.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

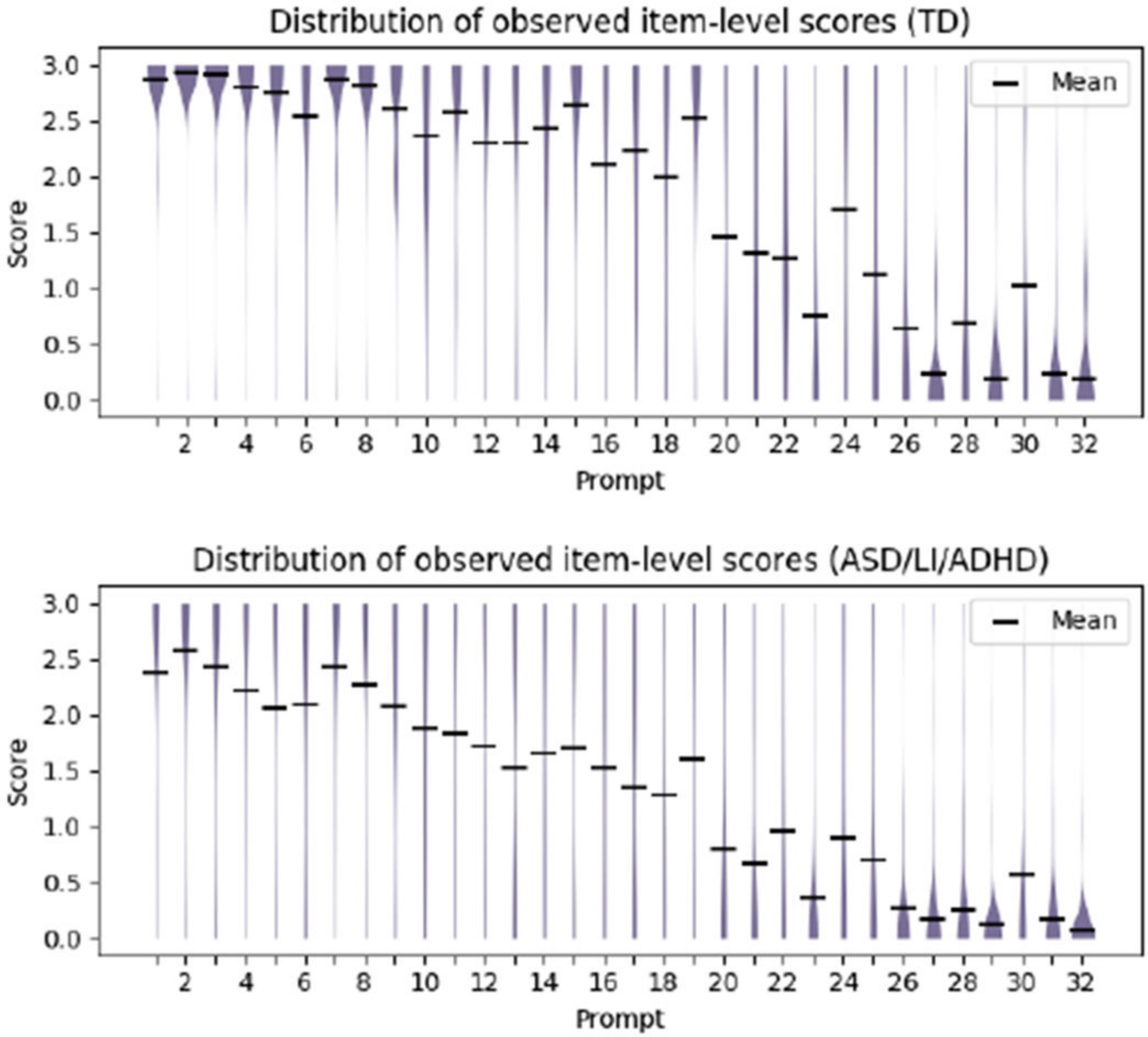

Fig. 1.

Distribution of true item-level scores across TD (top) and AD (bottom) participants' responses as a function of item number

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Prompt:	My	mother	is		a								doctor	who	works	in	a	city	hospital.
Response:	My	mother	is	is	a	is	a	(um)	is	a	city	(um)	doctor	that	works	in	a	city	hospital.
Expert labels:	C	C	C		C						A		C	S	C	C	C	C	C
Edit distance labels:	C	C	C	A	C	A	A		A	A	A		C	S	C	C	C	C	C

Fig. 2.

An example prompt and response. Since repairs and disfluencies do not count as errors in conventional scoring, this response would be scored with only two errors: one addition (A) and one substitution (S). However, although the “um”s are easily removed in preprocessing, the edit distance algorithm is not equipped to recognize repairs, and counts far more errors.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

Participant characteristics (Mean (SD))

Variable	TD (n=43)	ASD/LI/ADHD (n=61)
Age	6.51 (1.08)	6.97 (0.89)
Gender (% Female)	0.47 (0.50)	0.28 (0.45)
SRT score	59.32 (18.57)	42.63 (24.71)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE II

Performance using addition (A), substitution (S), omission (O), and correctly-aligned (C) measures.

Transcription	Predictor	MAE	PCC	# Feat.
-	Dummy	18.14	−1.00	128
Manual	Rules-based	7.31	0.98	128
ASR	Rules-based	8.63	0.94	128
Manual	Ridge (L2)	4.86	0.96	128
Manual	Linear SvR	**4.28**	0.97	128
ASR	Ridge (L2)	8.51	0.88	128
ASR	Linear SVR	**6.52**	0.92	128