INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Prediction of Speech Delay from Acoustic Measurements

Jason Lilley, Madhavi Ratnagiri, and H. Timothy Bunnell

Nemours Biomedical Research, Wilmington DE, USA

lilley@asel.udel.edu, ratnagiri@nemoursresearch.org, bunnell@nemoursresearch.org

Abstract

Speech delay is characterized by a difficulty with producing or perceiving the sounds of language in comparison to one's peers. It is a common problem in young children, occurring at a rate of about 5%. There are high rates of co-occurring problems with language, reading, learning, and social interactions, so intervention is needed for most. The Goldman-Fristoe Test of Articulation (GFTA) is a standardized tool for the assessment of consonant articulation in American English children. GFTA scores are normalized for age and can be used to help diagnose and assess speech delay. The GFTA was administered to 65 young children, a mixture of delayed children and controls. Their productions of the 39 GFTA words spoken in isolation were recorded and aligned to 3-state hidden Markov models. Seven measurements (state log likelihoods, state durations, and total duration) were extracted from each target segment in each word. From a subset of these measures, cross-validated statistical models were used to predict the children's GFTA scores and whether they were delayed. The measurements most useful for prediction came primarily from approximants /r, l/. An analysis of the predictors and discussion of the implications will be provided.

Index Terms: speech delay, hidden Markov models, automatic assessment, principal component analysis, ensemble linear regression

1. Introduction

Speech delay (SD) is characterized by a difficulty with producing and/or perceiving the sounds of one's native language in comparison to one's peers. In other words, a child with speech delay masters these sounds at a later-than-average age. Speech delay in young children is a very common problem, occurring at a rate of about 5% [1, 2, 3]. There are high rates of co-occurring problems with language, reading, learning, and social interactions, so intervention is needed for most [4, 5]. Symptoms and outcomes are heterogeneous: while many cases are resolved within a few years, others persist, and some result in diagnoses of more serious or specific syndrome such as dyslexia or specific language impairment [3, 6].

The Goldman-Fristoe Test of Articulation (GFTA) [7] is described as "a systematic means of assessing an individual's articulation of the consonant sounds of Standard American English" (pg. 1). Although it was not developed as a measure of SD in particular, the instrument produces a Standard Score that is normalized per the child's age and gender. This score can thus serve as a useful first approximation of the child's degree of SD.

The second edition of the GFTA (henceforth GFTA-2), used in this study, comprises 3 parts, only the first of which is used to produce the raw and standard scores. This part, called the Sounds-in-Words test, is designed to elicit utterances of 53 target words. The child is shown a series of 34 picture plates depicting everyday scenes, and is asked to name specific objects or actions in the scenes. The administrator does not model the target words, so the test is intended to be a measure of articulation in spontaneous speech. The administrator then grades particular target sounds – which are either single consonant segments or two-consonant clusters – in the words as pronounced either correctly or incorrectly. There are 77 targets (61 single consonants and 16 clusters) among the 53 words. The raw score, which is simply the number of targets produced incorrectly (0-77), is then converted to a Standard Score by looking up the child's raw score, gender, and age in a set of tables distributed with the test materials. The Standard Scores are normalized such that a score of 100 is average for the child's age and gender. Higher scores indicates higher-than-average performance, while lower scores indicate lower-than-average performance. The standard deviation is 15, though the authors urge caution in interpreting this, as scores are not normally distributed.

This paper describes the initial steps in a study of the genetics of speech delay. The hypothesis being examined in that study is whether different subtypes of SD, caused by different genotypes, give rise to different phenotypes that can be detected in specific measurements of the children's speech. If this is the case, then it may be expected that these measurements may also be used to distinguish children with SD as a whole from typically-developing (TD) children. This paper focuses on this second hypothesis by using specific measurements of GFTA-2 words to predict the child's GFTA-2 scores. We examine whether either the raw score or the standard score can be predicted accurately.

2. Data

This is a retrospective study, using data collected in a previous grant (NIH R21-DC007466-01, P.I. Bunnell).

2.1. Participants

Pairs of siblings from 34 families were recruited for the original genetics study, with the principal requirement being that at least one sibling must have been diagnosed with SD. In the case of one family, 3 siblings were entered into the genetics study, bringing the sum to 69, but 4 participants were dropped from this study, due to missing audio recordings. Hence the total number of participants in this study is 65. Participant ages ranged from 5;1 to 10;1 at the time of GFTA administration.

2.2. Data collection and scoring

An experienced speech-language pathologist administered the GFTA-2 to each participant, who wore a Sennheiser HMD 410 head-mounted microphone to capture their utterances. The audio was recorded at a 48-kHz sampling rate with 16-bit

http://dx.doi.org/10.21437/Interspeech.2017-1740

resolution. The target sounds in each word were later scored independently by two judges (members of the research team) by listening to the recordings through headphones. The judges subsequently discussed the sounds they disagreed on until a consensus score was reached. The total raw score (number of errors) was converted to a Standard Score for each child by using the conversion tables published with the test, indexed by age and gender. Raw scores ranged from 0 to 52, with a median of 10 and a mean of 13.3. Standard Scores ranged from 45 to 111, with a median of 87 and a mean of 86.1.

2.3. Acoustic measurement extraction

The hidden Markov model (HMM) systems were developed using tools in the Cambridge University Hidden Markov ToolKit (HTK) [8]. An unrelated set of child recordings was used to train models that were then aligned to the GFTA data.

2.3.1. Model training

A database of 18,531 multisyllabic single-word utterances recorded from 207 typically developing children between the ages of 6 and 8 was used to train phoneme HMMs using maximum-likelihood estimation. Single-Gaussian 3-state monophone HMMs were trained with feature vectors consisting of 39 Mel-frequency cepstral coefficients (13 static, 13 delta, and 13 acceleration). The vectors were calculated at a frame rate of 5 msec and a window size of 25 msec using a Hamming window.

2.3.2. Model alignment and measure extraction

The trained models were then force-aligned to the waveforms of the GFTA recordings. Note that, in general, a transcription of the word actually spoken, rather than the target word, was used to determine the model sequence to align to the recording, e.g. if the child produced *froggy* for the target word *frog*, then the models aligned to the recording were the 5 models corresponding to the 5 phonemes /frɑgi/ (plus initial and final silence models). However, in the case of the GFTA-2 target sounds, the models for the target phonemes were always used, even if the child's production sounded more like a different phoneme. For example, when /r/ was a target sound, the /r/ HMM was always used, even when the child's production sounded more like /w/.

After alignment, the following 7 measurements were extracted for each target phoneme in each recording:

- The total duration of the phone segment, as determined by the forced alignment;
- The fraction of that duration aligned to each of the model's 3 states; and
- The log-likelihood of alignment of each state to that portion of the wavefile, as calculated by the model, normalized by the state duration.

Note that even though a consonant cluster counts as a single "sound" for the purposes of GFTA-2 scoring, two 3-state HMMs are aligned to each cluster, and so 14 measures are extracted. There are 61 single consonant targets and 16 biphone consonant cluster targets in the GFTA-2, for a total of 93 (61+16*2) target phonemes. However, a few of the participants' recordings were either missing or unusable, so these words were excluded from further analysis for all participants. These exclusions reduced the total number of target sounds to 81, producing a total of 567 measures for each participant.

3. Experiments

3.1. Prediction of raw scores with PCA

The 567 raw measures from the 65 participants underwent a Principal Component Analysis (PCA), yielding 65 principal components. The first 7 principal components (cumulatively accounting for 27.8% of the variance) were used in a 65-fold (leave-one-out) cross-validation experiment. In turn, each of the participants' data was removed, and the first 7 principal components of the remaining data was used to build a linear model to predict the raw score. The fitted model was then used to predict the raw score of the held-out participant.

The Pearson correlation between the actual and predicted scores in this model was .776. When the participant's age (measured in months) was added as an eighth variable, the correlation improved to .818. Some predicted scores in this model, however, were negative, which would be impossible, since the raw score is a count and must be a natural number. Setting a floor on the model, by raising all negative scores to 0, improved the overall correlation to .819. This correlation is significantly different than the original correlation of .776 (*p* = .0473) [9, 10]. These results are illustrated in Figure 1. Adding more of the principal components to the models did not improve the correlation any further.

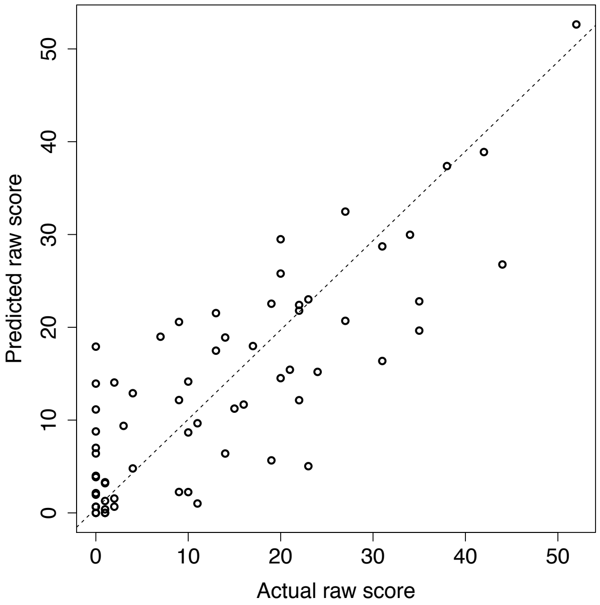

Figure 1: *Predicted versus actual raw GFTA scores, with fitted regression line, using cross-validated PCA models.*

The fitted regression line (pictured) has an intercept of 0.46 and a slope of 0.96, which is close to the ideal fit, with an intercept of 0 and slope of 1.

3.2. Prediction of standard scores with PCA

The 65-fold cross-validation experiment was repeated, this time with the participants' standard scores. As before, the 567 HMM-derived measures were initially used as the input to the PCA, and the first 7 components were used as the explanatory variables to the linear models. This yielded a cross-validated Pearson correlation of .632. Adding more components did not improve the correlation. Again we added age as another variable, but this time performance was most improved by including age as part of the PCA. Using 7 components again, the Pearson correlation increased to .635, although this increase was not significant (p = .2596) [9, 10]. This fit is shown in Figure 2.

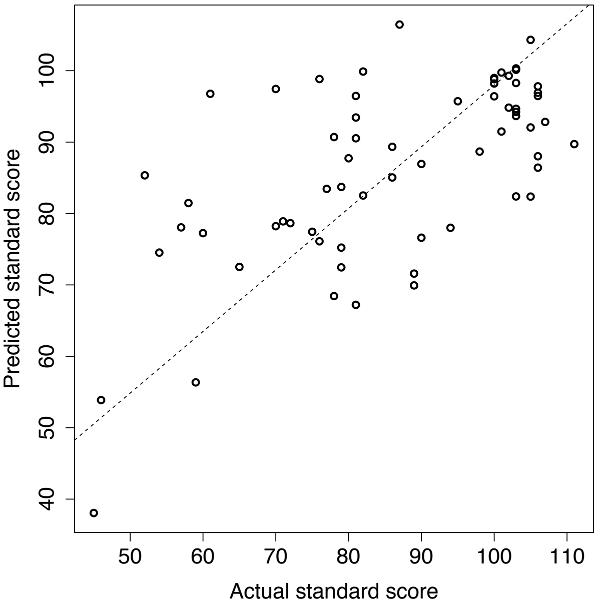

Figure 2: *Predicted versus actual standard GFTA scores, with fitted regression line, using cross-validated PCA models.*

The fitted regression line has an intercept of 11.73 and a slope of 0.86, which is considerably farther from the ideal fit than the model of the raw scores produced.

3.3. Prediction of standard scores with ensemble linear regression

While the PCA experiments described above show a good correlation with GFTA scores, they give little insight into which of the 567 HMM measures are most helpful in predicting the scores. The contributions of the measures to the principal components give one an understanding of the variance within the measures, but this variance may have little relation to the dependent variables. We examined the measures more closely by conducting a variation on ensemble linear regression [11].

3.3.1. Procedure

In this experiment we trained 500 linear regression models (henceforth ensemble models) on the GFTA standard score with the following algorithm. For each of 500 iterations, the participants were randomly divided into two groups: 33 for training the models, and 32 for testing. To choose the variables for each model, a greedy forward selection algorithm was used. Starting with a null set, in each round of construction, the variable that produced the highest increase in model performance on the test data was successively added to the model, until performance stopped increasing. The performance measure used was the root mean squared error (RMSE) between the actual and fitted standard scores of the test data.

The measures used for each ensemble model, and the order in which they were added, were recorded. These data give a robust idea of which measures are most useful for predicting the standard scores. The number of measures used for the models ranged from 20 to 31, with a mode of 27. Each of the 567 measures was selected by at least 3 models, while the most popular measure was selected 158 times. This measure was the log likelihood of the second state of the /p/ phoneme in the word *spoon*.

However, the number of times that a measure is selected overall is not necessarily indicative of its overall utility as a predictor, since so many predictors were used in each model. For example, a predictor selected as the last variable in 100 models is perhaps not as important as one selected 50 times as the first variable. To get a better picture, we conducted a further set of 270 cross-validation experiments. In each experiment, the first X (X=1, 2, 3, 4, 5, 10, 15, 20, all) predictors selected in each of the 500 ensemble models were pooled and ordered according to how often they were selected. Then the first Y (Y=1 to 30) highest-ranked predictors were used to build a set of leave-1-out cross-validated models (where one participant is held out, the model is built on the remaining data using the selected predictors, and the model is tested on the held-out participant).

Table 1: *The 17 measurements used in the final fitted models. LL: state log-likelihood (normalized by state duration). PD: proportion of duration aligned to the state.*

Word	Phone	Measure	Top 10 Count	Total Count
spoon	/p/	State 2 LL	158	158
glasses	/l/	State 2 PD	150	155
brush	/r/	State 1 PD	142	146
pajamas	/ʤ/	State 2 PD	141	141
pajamas	/ʤ/	State 3 PD	113	120
brush	/r/	State 2 PD	94	101
slide	/s/	State 2 LL	84	85
wagon	/g/	State 3 PD	82	90
glasses	/g/	State 1 PD	81	90
jumping	/ʤ/	State 3 LL	78	87
tree	/r/	Total Dur.	75	84
car	/r/	State 2 LL	68	70
glasses	/l/	State 2 LL	65	66
banana	/n/	State 2 PD	62	68
drum	/r/	State 1 PD	60	68
brush	/b/	State 1 LL	47	53
window	/w/	State 1 PD	47	54

The resulting correlations over all these experiments were recorded. The highest correlation, .893, was produced when Y=17 and X was either 10, 15, or 20. That is to say, this correlation was obtained when either the first 10, 15 or 20 predictors from the ensemble models were pooled and ranked, and then the first set of 17 predictors was chosen from the ranked set to produce the final cross-validated models. This is because the 17 predictors were identical in each of the three cases. This set of 17 predictors is listed in Table 1.

The fit of these 17 predictors to the standard scores is illustrated in Figure 3. The fitted regression line has an intercept of 4.65 and a slope of 0.94, which is closer to the ideal fit than in Figure 2.

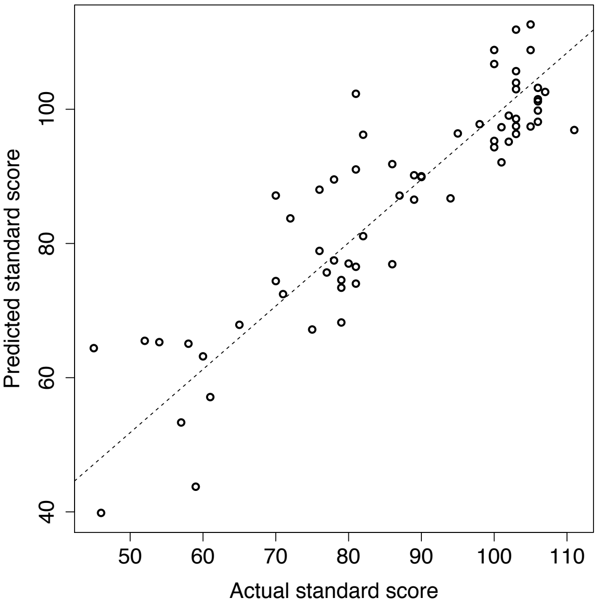

Figure 3: *Predicted versus raw GFTA scores, with fitted regression line, using 17 predictors chosen from ensemble models.*

3.3.2. Discussion

Some caution must be taken when interpreting Table 1. This experiment was not completely cross-validated, in that all of the data was used as part of the procedure to identify these 17 measures. Although the final cross-validation experiment described suggests that the measures are somewhat robust, it is possible that the data are over-fitted, and these measures may not generalize to other data.

With that in mind, it is interesting to consider the measures selected. We note that 7 of the 17 measures are derived from the liquid approximants /r,l/. This is not unexpected, for two reasons. The first reason is that children often have trouble with these phonemes, and master them late relative to other single consonants [7]. The second reason is that, probably due to the first reason, these phonemes occur as targets often in the GFTA words, both in isolation (3 times each) and as part of target clusters (6 times each). These 18 appearances account for 23% of the GFTA targets. Perhaps more surprising is that the affricate /ʤ/ occurs 3 times in the list, but this is another phoneme that children master relatively late; the GFTA manual (p. 9) states that the proportion of children in the test standardization study who mastered the phoneme did not reach 85% until the age of 5 (equal to /l/; /r/ was mastered by age 6).

Another angle to examine is which sorts of measurements occur on the list. The proportion of the phoneme segment assigned to each state is the most common measure on the list, although log likelihoods, which are perhaps the most direct measurement of the deviance of the production, are also common. Total phone duration only occurs once (for the /r/ in *tree*), even though children with articulation difficulties sometimes produce rhotic segments of exaggerated length [12].

4. Conclusions

We find that positive initial steps have been taken, using the HMM-derived measures described in this paper, to produce a model that can predict a child's GFTA score, and hence indirectly measure the degree of their speech delay. While the modest fit is a fine starting point, there is certainly much room for improvement. In the future, we will examine other types of measurements (from HMMs or otherwise) that may produce a better fit. One option is to use HMMs with skippable states, or even multiple states placed in a parallel, mutually exclusive configuration; the particular series of states that are aligned to the signal may indicate an aberrant pronunciation. Another possibility is to use some sort of confidence measure, derived from the fit of signal to the model of the expected phoneme relative to models of other phonemes.

Recall that this study was conducted under the assumption that in order to determine whether a set of acoustic measurements can distinguish speakers with different types of speech delay, it would also be necessary to demonstrate that some set of measurements can distinguish any speech-delayed speaker from the set of typically-developing children. This assumption, however, may turn out to be false. It may not be necessary to find a unique set of measurements that all typically-developing children have in common, or that all speech-delayed children have in common. Instead, it is perhaps possible that there is a distinct set of measurements unique to each genetically distinct subset of speech-delayed speakers. If all of these subsets can be identified along with sufficiently diagnostic measures, then it will not be necessary to construct a model for typically-developing children after all; they will simply be identified by testing negative for all the speech-delay subtypes.

5. Acknowledgements

The research reported in this paper was supported by the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number U54-GM104941 (PI: Binder-Macleod). The original data used in this study was produced by a prior NIH grant (R21-DC007466-01, P.I. Bunnell).

6. References

[1] National Institute on Deafness and Other Communication Disorders (NIDCD), "Strategic Plan Fiscal Year 2006-2008," 2005.

[2] American Speech-Language-Hearing Association (ASHA), "Schools Survey report: SLP caseload characteristics trends 1995-2010," 2010.

[3] L. D. Shriberg, J.B. Tomblin, and J. L. McSweeny, "Prevalence of speech delay in 6-year-old children and comorbidity with language impairment," *Journal of Speech, Language, and Hearing Research,* vol. 42, no. 6, pp. 1461-1481, 1999.

[4] J. A. Gierut, "Treatment efficacy: Functional phonological disorders in children," *Journal of Speech, Language, and Hearing Research,* vol. 41, no. 1, pp. S85-100, 1998.

[5] B. Dodd, A. Holm, Z. Hua et al., "English phonology: Acquisition and disorder," *Phonological development and disorders in children: A multilingual perspective,* Z. Hua and B. Dodd, eds., pp. 25-55. Tonawanda, NY: Multilingual Matters Ltd, 2006.

[6] L. D. Shriberg, J. Kwiatkowski, S. Best et al., "Characteristics of children with phonologic disorders of unknown origin," *Journal*

of Speech and Hearing Disorders, vol. 51, no. 2, pp. 140-61, 1986.

[7] R. Goldman and M. Fristoe, *Goldman-Fristoe Test of Articulation*, 2nd ed. Circle Pines, MN: American Guidance Service, 2000.

[8] S. J. Young, "The HTK Hidden Markov Toolkit: Design and philosophy," Department of English, Cambridge University, 1993.

[9] J. B. Hittner, K. May, and N. C. Silver, "A Monte Carlo evaluation of tests for comparing dependent correlations," *Journal of General Psychology*, vol. 130, pp. 149-168, 2003. pmid: 12773018

[10] B. Diedenhofen and J. Musch, "cocor: A comprehensive solution for the statistical comparison of correlations." *PLoS ONE*, vol. 10, no. 4, e0121945. doi:10.1371/journal.pone.0121945

[11] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, pp. 123-140, 1996.

[12] H. T. Bunnell and J. Polikoff, "Acoustic characterization of children with speech delay," in *INTERSPEECH 2006 – 7th Annual Conference of the International Speech Communication Association, September 17-21, Pittsburgh, Pennsylvania, Proceedings*, 2006, pp. 989-992.