Computer-Assisted Syllable Complexity Analysis of Continuous Speech as a Measure of Child Speech Disorders

Marisha Speights Atkins[1], Suzanne E. Boyce[2], Joel MacAuslan[3], Noah Silbert[2]

[1] Auburn University; [2] University of Cincinnati; [3] Speech Technology and Applied Research Corp.
speighma@auburn.edu; boycese@ucmail.uc.edu; JoelM@STARAnalyticalServices.com; silbernh@ucmail.uc.edu

ABSTRACT

A common indicator of speech production disorders in children is a reduced ability to articulate complex syllables. Clinical studies of syllabic complexity of child speech have traditionally relied on phonetic transcription by trained listeners to characterize deviations in phonotatic structure. The labor-intensive nature of transcribing, segmenting, labeling, and hand-counting syllables has limited clinical analysis of large samples of continuous speech. In this paper, we discuss the use of a computer-assisted method, Automatic Syllabic Cluster Analysis, for broad transcription, segmentation, and counting syllabic units as a means for fast analysis of differences in speech precision when comparing children with and without speech-related disorders. Findings show that the number of syllabic clusters per utterance is a significant indicator of speech disorder.

Keywords: continuous speech, automated analysis, syllables, child speech

1. INTRODUCTION

In the course of development, children become more and more facile at voluntary coordination of the motoric movements necessary for utterance of complex syllables [14,15,35]. The mastery of complex syllables has been shown to be a powerful predictor of later communication skills [10,23,25]. Children with delayed speech acquisition do not show this same developmental trend and deviations in syllable acquisition may serve as diagnostic marker of future speech delay [7,8,20]

Historically researchers have been challenged to find ways to quantify maturation of speech production in young children, particularly in continuous speech [18,30]. Conventional approaches have typically involved careful transcription of child speech by researchers with specialized training. The time and labor intensive nature of such studies has necessarily limited the number and comprehensiveness of research studies [9,24]. This is a particular problem in studies of children with speech disorders that affect intelligibility [9].
Further, the acknowledged best way to characterize child speech is from longer, spontaneous productions, but the need for transcription and analysis of this large volume of data typically limits standardized testing to shorter, more controlled utterances [20,25]. To address this limitation, researchers have turned to automatized methods to quantify patterns in child vocalizations [6,17,24,33,35,36].

Automated approaches using acoustic features have been used in a number of studies [1-5, 24, 32, 35]. Another common approach is to employ ASR (automatic speech recognition) as an alternative to hand transcription [6, 36, 37]. ASR however, requires training of a large database of words and has been challenged by child speech [26]. In the case of very young children and children with speech impairments that affect intelligibility, although speech may be perceptually difficult to understand by humans, phonotatic patterns are still able to be detected [9-10].

To characterize difference in syllable complexity in children with and without speech disorders, we utilize an automated approach that is designed to detect syllabic units without reliance on word identification. The SpeechMark® landmark analysis system segments syllables using acoustic landmarks. Acoustic landmarks do not depend on the intelligibility of phonemes to characterize speech but focus on detecting change in acoustic patterns based on articulatory movements in the vocal tract and articulatory timing [2,3] These acoustic events that occur based on changes of the articulatory shape of the oral cavity are called landmarks. Unlike approaches based on ASR, this method does not depend on identification of specific words or speech sounds in the signal. Instead, the pattern of landmarks provides a method for

tracking articulatory patterns held in common across different syllables, words or sentences. Thus, landmark-based tools are particularly well- suited for analysis of non-lexical (independent of phonemic identification) articulatory differences in the way a word or syllable is produced [7,8,27].

1.1 SpeechMark® Landmark Analysis System

The SpeechMark® system is based upon the prior work of Stevens et al. [27-29], Liu [21], and Howitt [12] Abrupt landmarks occur at points where abrupt changes in the amplitude of several frequency bands reach predetermined thresholds for detection. Landmark detection occurs by first computing a spectrogram with a 6ms Hanning window every 1ms. The spectrogram is then divided into the six frequency bands, ranging 0.0 −0.4, 0.8−1.5, 1.2−2.0, 2.0−3.5, 3.5−5.0, and 5.0−8.0 kHz [12,21]. The algorithm localizes moments where abrupt changes and peaks in energy occur. An "abrupt" change occurs when power increases/decreases at the minimum of 6 dB simultaneously in the finely and coarsely smoothed contours [13]. When the energy change is not sufficient to meet the threshold the landmark will not be detected.

Abrupt glottal and oral landmark types used in this study, and their mnemonic labels:

- g: glottis. Marks the beginning (+g) and end (-g) of sustained laryngeal motion near a segment of sustained periodicity.
- p: periodicity: Marks the beginning (+p) and end (-p) of sustained periodicity of an appropriate period.
- s: syllabicity: Marks sonorant consonantal releases (+s) and closures (-s) in a voiced segment
- b: burst: Marks frication onsets or affricate/stop bursts (+b) and the point where aspiration or frication ends (-b) in an unvoiced segment
- f: unvoiced frication onset (+f) and offset (-f) of simultaneous power increases/decreases or decreases/increases of high frequencies/low frequencies respectively
- v: voiced frication onset and offset of simultaneous power increases/decreases (+v) or decreases/increases (-v) of high frequencies/low frequencies respectively

Non-abrupt landmark

V: Vowel: Marks a time point corresponding to local maximum harmonic power

F: Continuing frication: marks a time of maximally well-developed air turbulence.

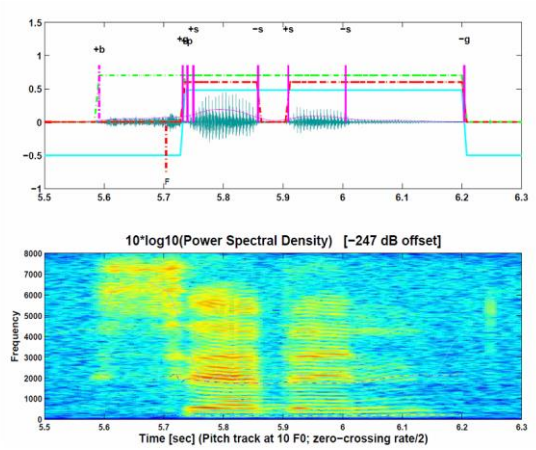
1.2. Syllabic Cluster Analysis

SpeechMark's extension of the landmark system, *Automatic Syllabic Cluster Analysis*, further groups the landmarks into syllabic units. Syllabic Clustering occurs after landmarks have been detected in the signal. Consecutive abrupt oral and glottal landmarks are grouped around non-abrupt vowel landmarks into clusters that approximately match the shape of phonotactically well-formed linguistic syllables. Syllables with a simple structure will show fewer landmarks, while syllables with a complex structure will show more landmarks and more unusual patterns of landmarks[3].

The method detects differences in articulatory precision as a difference in how many of the landmarks characteristic of the canonical form are present, and in what pattern. Recall that the detection of abrupt landmarks is based on thresholding---when a syllable is produced with articulatory movements that are strong and timed appropriately, the acoustic signatures will meet the threshold, and the full set of landmarks will be detected. When the same syllable is spoken with fewer extreme articulatory movements in a shorter period of time, the landmark thresholds will not be reached, and fewer landmarks will be detected. Thus, different syllabic cluster groupings will be detected when (1) a string of intended syllables is produced in its canonical form (CCVC), (2) in a less complex form (CVC), or (3) a more *lenited, (i.e.* softened consonant) form (/t/ becomes /ɵ /, /s/ becomes /f/, etc. For example, the sequences "aah", "bah", and "bat" consist of V, CV, and CVC syllables. If produced canonically, they would be distinguished from one another as +g-g, +b+g+s-g, and +b+g+s-g+b-b. Note that, because of the non-lexical aspect of the SpeechMark system, many words with similar syllabic shape will show up with identical sequences of landmarks; for example, if produced in canonical form, the words "backed" and "that's" will both appear as +g+s-g+b-b. If "backed" is produced with a more weakly produced final stop consonant—i.e. more like "back" with an unreleased final "k"—it will appear as +g+s-g.

Figure 1: Landmarks and Syllabic Clusters. The segment shows the word *seven*, together with the landmarks found (*magenta*, solid lines for high strength, otherwise dotted). Dentiles over the waveform mark the syllabic clusters of the two syllables (*dashed red*, 5.72-5.85s and 5.91-6.20s). Other

dentiles mark the voiced segment (*solid cyan*, 5.72-6.20s), which extends through both syllables, and the utterance cluster (*dashed green*, *top*, either 5.60-6.20s or 5.72-6.20s, depending on the strength threshold of landmarks to be included). The example also shows a peak-type landmark (5.70s).



1.3 Toward Automated Speech Disorder Detection

Both landmark patterns in general, and Syllabic Clusters in particular, have been shown to correlate with changes in articulatory precision in normal and disordered speakers [2,3,]. Studies have shown that syllabic cluster patterns are significantly different in clear vs. conversational speech, and can distinguish more- vs. less-intelligible speakers [4]. Adult speakers with disorders show different patterns of landmarks and syllabic clusters relative to typical speakers [3,5,13].

As noted above, studies using transcription and syllabic coding have shown that children become more adept at producing complex syllables as a function of age and practice. Further, children with speech disorders lag behind their peers in this skill [14,31,35]. Accordingly, we expect to find that children with speech sound disorders, will produce fewer syllabic clusters, and that the clusters themselves will be simpler. In the case of child speech, complex syllables that an adult might produce as a syllabic cluster with multiple landmarks (e.g. "string") may be produced as a simpler syllable with fewer landmarks ("sring", "ring"), or even be broken into two syllables of simpler structure ("siring"). Depending on the degree to which the transcriber is aware of small phonetic differences, "siring" might be transcribed as /sərɪŋ/ or /strɪŋ/. In this paper we apply the Automatic Syllabic Cluster Analysis to differentiate between children who have typically developing speech production and those with speech production disorders.

2. METHODS

2.1. Participants

The speech of 37 children ages 3-5 was analyzed using Syllabic Cluster Analysis to measure differences between groups. Of the children, 27 were typically developing and 10 were diagnosed with speech sound disorder. Speech was recorded in either of two locations—in a quiet room at a community preschool or a university clinic. Children were screened for normal hearing using the criterion of sound detection at 20 dB HL for pure tones at 500, 1000, 2000, and 4000 Hz. Speech and language were assessed using the Clinical Assessment of Articulation and Phonology 2nd edition [34] and the Clinical Evaluation of Language Fundamental-preschool 2nd Edition [38]. Children with standard scores of 80 or greater were considered to be typically developing. Children with lower scores were considered to fit the diagnosis of Speech Sound Disorder.

2.2 Recordings

Continuous speech samples were elicited using a child story book with repetitive sentences for each subject (Brown Bear, Brown Bear, What do You See [16]). Recordings were collected using a Shure wireless system with a unidirectional, cardioid lavalier microphone and receiver connected to laptop computers. Speech was recorded directly using WaveSurfer (www.speech.kth.se/wavesurfer/). The Shure body pack transmitter and microphone was worn by the child on a well-fitted vest. Samples were digitally processed at a sampling rate of 22K and 24bit depth. A total of 1172 sentences from 37 child speakers were recorded for analysis.

2.3 Instrumentation

Syllable complexity was measured using the Syllabic Cluster Analysis algorithm in the SpeechMark ® Matlab toolbox. The following parameters were analyzed: number of (1) landmarks (LMs), (2) syllabic clusters (SCs), (3) landmarks per syllabic cluster (LM/SC), and (4) syllabic clusters per utterance (SC/Utts). LM/SC and SC/Utts were used in logistic regression models as predictors of disordered group status.

3. RESULTS

Mixed-effects logistic regression models were fit to analyze how well LM/SC and SC/Utts

predict disordered group status. Mixed models are useful for analyzing multiple observations within subjects, providing flexibility in modeling expected values, as well as between-subject differences simultaneously [11]. The mixed-effect and standard models yielded equivalent results.

Table 1: Logistic regression model fits for landmarks per syllable (top row) and syllable clusters per utterance (bottom row). Slope estimates (first column), standard errors (second column), z values (third column), and p values (fourth column).

	Estimate	SE	z value	Pr(>|z|)
LM/SC	0.01	0.05	0.17	0.85
SC/Utts	-0.21	0.031	-6.85	<.001

The slope estimate and standard error for LM/SC indicates that LM/SC is a very poor predictor of disordered group status. On the other hand, the slope estimate and standard error for SC/Utts indicates that SC/Utts is a useful predictor of disordered group status. The slope estimate of -0.21 indicates that for each unit increase in one SC/Utts, the probability of being in the TD group decreases by approximately 0.05.

Figure 2. Landmarks per syllable cluster (LM/SC) and disordered group status. The x-axis indicates group status (TD = Typically Developing; D = Disordered), and the y-axis indicates LM/SC. The boxplots indicate the median (thick horizontal line) and interquartile region (lower and upper box limits), with the notch indicating an approximate 95% confidence interval for the median. The whiskers indicate 1.5 times the interquartile range, with dots indicating data outside this range.

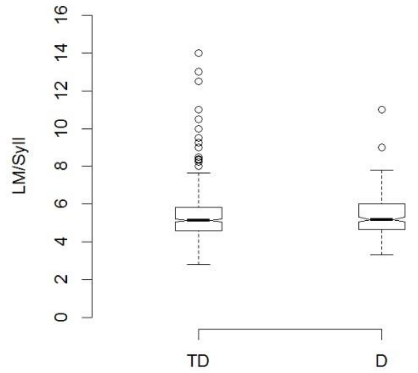
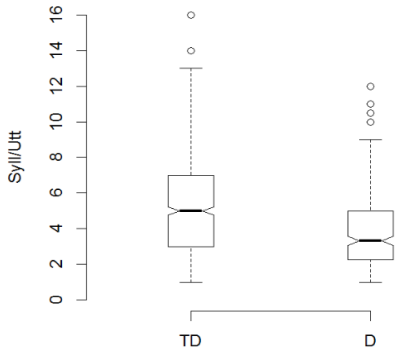


Figure 3. Syllable clusters per utterance (SC/Utts) and disordered group status. The x-axis indicates group status (TD = Typically Developing; D = Disordered), and the y-axis indicates SC/Utts. The boxplots indicate the median (thick horizontal line) and interquartile region (lower and upper box limits), with the notch indicating an approximate 95% confidence interval for the median. The whiskers indicate 1.5 times the interquartile range, with dots indicating data outside this range.



4. DISCUSSION

In this study, we applied the Automatic Syllabic Cluster Analysis approach as a computerized method for analyzing continuous speech samples recorded by preschool aged children. Syllabic complexity was measured as a means of detecting differences between talkers who were identified to have typically developing speech production and those with speech production disorders. Syllabic clusters per utterance was found to be a significant predictor of disordered speech in running speech samples. One of the challenges in measuring the speech of older children with speech disorders is that errors may overlap with typical developmental speech errors in younger children.

5. CONCLUSIONS

Automated Syllabic Cluster detection is useful for detecting differences in speech production utilizing a landmark-based technology. This approach could serve 1) as a computer-assisted approach to measuring syllable complexity when analyzing speech spoken by preschool-age children and particularly those with decreased intelligibility due to speech disorders and, 2) a means for early identification of children who are developing on abnormal trajectories using continuous speech samples.

6. ACKNOWLEDGEMENTS

This work was sponsored by the National Institutes of Health grants 5R44DC010104-04, and 3R44DC010104-03S1.

At present, SpeechMark® is available without cost to researchers (www.speechmrk.com)

7. REFERENCES

1) Boyce, S. E., Balvalli, S. N., MacAuslan, J., Martin, D., & Clark, J. (2011). Objective Data on Clear Speech: Does it Help in Training Audiology Students. *Poster at AudiologyNow2011*.
2) Boyce S, Fell HJ, MacAuslan J. SpeechMark: Landmark Detection Tool for Speech Analysis. Paper presented at: INTERSPEECH2012.
3) Boyce, S., Fell, H., Wilde, L., & MacAuslan, J. (2011). Automated Tools for Identifying Syllabic Landmark Clusters that Reflect Changes in Articulation. In *Models and Analysis of Vocal Emissions for Biomedical Applications* (pp. 63–66). Firenze: Firenze University Press. Chenausky, K., MacAuslan, J., & Goldhor, R. (2011). Acoustic Analysis of PD Speech. *Parkinson's Disease*, *2011*, 1–13.
4) Boyce, S., Krause, J., Hamilton, S., Smiljanic, R., Bradlow, A. R., Rivera-Campos, A., & MacAuslan, J. (2013, June). Using landmark detection to measure effective clear speech. In *Proceedings of Meetings on Acoustics*(Vol. 19, No. 1, p. 060129). Acoustical Society of America.
5) Boyce, S., Ishikawa, K., & Speights, M. (2014). Intelligibility in dysphonic speech: Landmark-based measures. *The Journal of the Acoustical Society of America*, *135*(4), 2292-2292.
6) Chen, Y. J. (2011). Identification of articulation error patterns using a novel dependence network. *IEEE Transactions on Biomedical Engineering*, *58*(11), 3061-3068.
7) Fell, H. J., MacAuslan, J., Ferrier, L. J., & Chenausky, K. (1999). Automatic babble recognition for early detection of speech related disorders. *Behaviour & Information Technology*, *18*(1), 56-63.
8) Fell, H., MacAuslan, J., Ferrier, L. J., & Worst, S., Chenausky, K., (2002). *Vocalization Age as a Clinical Tool*. Proceeding from ICSLP: International Conference on Speech Processing.
9) Flipsen Jr, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical linguistics & phonetics*, *20*(4), 303-312.
10) Flipsen Jr, P. (2006). Syllables per word in typical and delayed speech acquisition. *Clinical linguistics & phonetics*, *20*(4), 293-301.
11) Gelman & Hill (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press
12) Howitt, A. W., *Automatic Syllable Detection for Vowel Landmarks*, doctoral thesis M.I.T., Cambridge, MA. 2000.
13) Ishikawa, K., Rao, M. B., MacAuslan, J., & Boyce, S. (2019). Application of a Landmark-Based Method for Acoustic Analysis of Dysphonic Speech. *Journal of Voice*.
14) Masso, S., McLeod, S., Baker, E., & McCormack, J. (2016). Polysyllable productions in preschool children with speech sound disorders: Error categories and the Framework of Polysyllable Maturity. *International journal of speech-language pathology*, *18*(3), 272-287.
15) MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and brain sciences*, *21*(04), 499-511.
16) Martin, B., & Carle, E. (1984). *Brown bear, brown bear*. Puffin books.
17) Middag, C., Martens, J. P., Van Nuffelen, G., & De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, *2009*(1), 629030.
18) Kent, R. D. (1996). Hearing and believing some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*(3), 7- 23.
19) Klein, H. B. (1981). Productive strategies for the pronunciation of early polysyllabic lexical items. Journal of Speech, Language, and Hearing Research, 24(3), 389-405.
20) Klein, H. B., & Liu-Shea, M. (2009). Between-word simplification patterns in the continuous speech of children with speech sound disorders. Language, speech, and hearing services in schools, 40(1), 17-30.
21) Liu S. A. (1994)Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*.;100(5):3417-3430.
22) Oller, D. K. (2000). *The emergence of the speech capacity*. Psychology Press.
23) Oller, D. K., Eilers, R. E., Neal, A. R., & Schwartz, H. K. (1999). Precursors to speech in infancy: the prediction of speech and language disorders. *Journal of communication disorders*, *32*(4), 223- 245.
24) Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354-13359.
25) Paul, R., & Jennings, P. (1992). Phonological behavior in toddlers with slow expressive language development. *Journal of Speech, Language, and Hearing Research*, *35*(1), 99-107.
26) Shivakumar, P. G., & Georgiou, P. (2018). Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. *arXiv preprint arXiv:1805.03322*.
27) Stevens, K.N. 1992. Lexical access from features. *Speech Communication Group Working Papers*, Volume VIII, Research Laboratory of Electronics,
28) Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872-1891
29) Stevens, K.N., Manuel, S., Shattuck-Hufnagel and Liu, S. 1992. Implementation of a model for lexical access based on features. *Proc. Int'l. Conf. Spoken Language Processing*, Banff, Alberta, **1**, 499-502.
30) Stoel-Gammon, C. (2001). Transcibing the Speech of Young Children.*Topics in language disorders*, *21*(4), 12-21.
31) Stoel-Gammon, C. (2010). The word complexity measure: Description and application to developmental phonology and disorders. *Clinical linguistics & phonetics*, *24*(4-5), 271-28
32) Speights, M., Boyce, S., MacAuslan, J., Ishikawa, K., Fell, H., and Ungruhe, J. (May, 2015). *Measurement of child speech complexity using acoustic landmark detection.* Acoustical Society of America. Pittsburgh, PA.
33) Speights, M. L., Ishikawa, K., Boyce, S, MacAuslan, J., Fell, H., Ungruhe, J., Longpre, K. (2015, November). *Automatic Syllabic Cluster of Children's Speech Data to Indentify Speech-Disorders*, American Speech Language and Hearing Association National Convention Denver, CO.
34) Secord, W., Donohue, J. A. S., & Super Duper Publications (Firm). (2002). *CAAP: Clinical Assessment of Articulation and Phonology*. Greenville, S.C: Super Duper Publications
35) Vick, J. C., Campbell, T. F., Shriberg, L. D., Green, J. R., Truemper, K., Rusiewicz, H. L., & Moore, C. A. (2014). Data-driven subclassification of speech sound disorders in preschool children. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2033-2050.
36) Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated Analysis of Child Phonetic Production Using Naturalistic Recordings. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1638-1650
MIT, 119-144.
37) Ward, L., Stefani, A., Smith, D., Duenser, A., Freyne, J., Dodd, B., & Morgan, A. (2016). Automated screening of speech development issues in children by identifying phonological error patterns. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)* (pp. 2661-2665).
38) Wiig, E. H., Secord, W., & Semel, E. M. (2004). *CELF preschool 2: clinical evaluation of language fundamentals preschool*. Pearson/PsychCorp.