International Journal of Technology Enhanced Learning

ISSN online: 1753-5263 - ISSN print: 1753-5255
https://www.inderscience.com/ijtel

Prosodic characterisation of children's Filipino read speech for oral reading fluency assessment

Francis D. Dimzon, Ronald M. Pascual

DOI: 10.1504/IJTEL.2023.10052486

Article History:

Received:	18 October 2021
Accepted:	23 November 2021
Published online:	22 December 2022

Prosodic characterisation of children's Filipino read speech for oral reading fluency assessment

Francis D. Dimzon* and Ronald M. Pascual

College of Computer Studies,
De La Salle University,
Manila, Philippines
Email: francis_dimzon@dlsu.edu.ph
Email: ronald.pascual@dlsu.edu.ph
*Corresponding author

Abstract: This paper explores the extraction and analysis of prosodic features in children's Filipino speech for application in automated oral reading fluency assessment. Automatic syllabication was optimised in the context of children's Filipino read speech. Using the Children Filipino Speech Corpus, prosodic features were automatically extracted which were then classified according to human rater assessment of fluency. Analysis of variance showed that speech and articulation rates, pauses, syllable duration, and pitch can be used to classify children's oral reading fluency in Filipino into three levels, namely, independent, instructional and frustration. Using machine learning classification methods, fivefold cross-validation showed that speech rate, articulation rate and number of pauses can be used to predict oral reading fluency at 92%, 85% and 76% accuracy for 2, 3 and 4 levels of fluency classification, respectively. Pitch and syllable duration patterns were also characterised for the assessment of phrasing and expression between fluent and non-fluent readers.

Keywords: oral reading fluency assessment; prosody; Filipino language; children's read speech.

Reference to this paper should be made as follows: Dimzon, F.D. and Pascual, R.M. (2023) 'Prosodic characterisation of children's Filipino read speech for oral reading fluency assessment', *Int. J. Technology Enhanced Learning*, Vol. 15, No. 1, pp.74–94.

Biographical notes: Francis D. Dimzon obtained his BSc in Applied Mathematics from the University of the Philippines Visayas in 1994 and MSc in Computer Science from the University of the Philippines Los Baños in 2000. He is currently a PhD student in De La Salle University, Manila, Philippines. His research interests include speech processing and machine learning.

Ronald M. Pascual is with the Computer Technology Department, College of Computer Studies, De La Salle University. He received his PhD degree in Electrical and Electronics Engineering from the University of the Philippines Diliman (UPD), his MS degree in Electronics and Communications Engineering from De La Salle University (DLSU) Manila, and his BS degree in Electronics and Communications Engineering from Pamantasan ng Lungsod ng Maynila (PLM). His research interests include audio and speech processing, children's speech technology, computer-aided language learning systems, computational linguistics, and digital signal processing.

This paper is a revised and expanded version of a paper entitled 'Computational Prosodic Features Analysis of Children's Filipino Speech for Automated Oral Reading Fluency Assessment' presented at the '20th Philippine Computing Science Congress', University of the Cordilleras, Baguio City, Philippines, 19–21 March 2020.

1 Introduction

An alarming concern in reading and comprehension skills among children in the Philippines points to the poor performance of Filipino students in the 2018 ranking results in Mathematics, Science, and Reading released by the Organisation for Economic Cooperation and Development (OECD)-Programme for International Student Assessment (PISA) (OECD, 2019). Of the 79 participating countries, the Philippines ranks last in Reading. The Philippine Informal Reading Inventory (Phil-IRI) test results also showed poor performance in reading comprehension CNN Philippines, 2020; Jaucian, 2020; Mocon-Ciriaco, 2020). Other compelling issues concern many students who are deficient in reading, numeracy, and comprehension (Bicol Standard, 2019). There are non-readers and frustration readers in high school (Albano Jr., 2019a; David et al., 2019; GMA TV7, 2018). Furthermore, the last five years has shown a consistently low performance of Grade 6 students in the National Achievement Test (NAT) (Albano Jr., 2019b; Hernando-Malipot, 2019). The average mean percentage score (MPS) for year 2018 declined to 37.44. This score is the lowest since the start of standardised evaluation administered by the Department of Education (DepEd). The 2019 Trends in International Mathematics and Science Study (TIMSS) revealed that the Philippines scored "significantly lower" than any other country that participated in Grade 4 mathematics and science assessments and ranked last among all 58 participating countries for both subject assessments (Magsambol, 2020; National Center for Education Statistics, 2020). Lastly, in the Southeast Asia Primary Learning Metrics of 2019 (SEA-PLM, 2019), among the six ASEAN countries, the Philippines ranked fourth or fifth, never higher in the rankings (SEA-PLM, 2019). SEA-PLM measures the learning outcomes of children enrolled in Grade 5.

1.1 Background

Reading involves two major processes – word decoding and comprehension (Rasinski, 2004). Educators assess reading competencies by relating reading capability to the ability to translate text orally (Fuchs et al., 2001). The ability to read aloud the text correctly (accuracy), with minimal use of attentional resources (automaticity), and to appropriately use phrasing, expression, volume, smoothness and pace (prosody) are what constitute oral reading fluency (ORF) (Rasinski, 2004). Thus, ORF assessments include assessing accuracy, automaticity and prosody. Accuracy can be measured by counting the number of words correctly read. Automaticity can be measured by counting the number of words correctly read per unit of time. Prosody is a valid yet subjective component of oral reading fluency (Rasinski, 2004).

Prosodic reading and comprehension are directly related to each other (Dowhower, 1987; Dowhower, 1991; Schreiber, 1991; Schreiber and Read, 1980). When an improved level of automaticity occurs, the reader constructs meaning to the text by expressive reading. Assessing prosody uses multidimensional rubrics (Zutell and Rasinski, 1991)

which cover expression and volume, phrasing, smoothness, and pace and involves subjective and qualitative scoring against specified criteria. Aside from English language, other studies would also use multidimensional rubrics to assess fluency, like in French (Godde et al., 2017) or in Spanish (Larsen, 2016). Measurement of prosody tends to be less reliable as human raters' consistency in scoring was average or low (Godde et al., 2017; Haskins and Aleccia, 2014; Mostow and Duong, 2009).

1.2 Automated oral reading fluency assessments

Automated oral reading fluency assessments were made possible by advancements in speech recognition. Some systems include Technology-Based Assessment of Language and Literacy (TBALL) (Alwan et al., 2007), LISTEN (Beck et al., 2004; Mostow and Duong, 2009), FLORA (Bolaños et al., 2011), and MAP ReadingFluency (Schaffhauser, 2019; NWEA, 2019). The TBALL project is geared towards automatic English literacy skills assessment of children, some of whom are native speakers of English while others speak English as a second language. It uses a specially designed language model to detect lexical disfluency in kindergarten students who read isolated words. However, lexical disfluencies are not sufficient in measuring oral reading fluency. The approach of LISTEN in evaluating reading proficiency among students is the use data collected from a student's interaction with a computer tutor. It measures prosody in oral reading by correlating that of children to adults in the assumption that adults' prosody is the correct one. However, the adult speech becomes incomparable to that of a child when a child makes a substitution or omission miscue. FLORA uses lexical features in rating the overall literacy of children in Grades 1 to 4. As its metric of oral reading fluency, it accounts the number of words correctly read per minute (WCPM). Nonetheless, FLORA produces more variability for individual student's scores when compared to expert human raters. Hence, there is a need to improve upon the methods used in FLORA. MAP ReadingFluency is an online tool that assesses oral reading fluency for speakers of English and Spanish. It uses EduSpeak (SRI International, 2019) speech recognition toolkit. Eduspeak, however, has no speech corpora for children in Filipino. There exists a children's Filipino speech corpus (CFSC) (Pascual and Guevara, 2012) but CFSC is not yet available for public use. One study considers prosodic features like emphasis and phrasing and decoding errors for ORF assessments (Sabu and Rao, 2018). However, performance of the automated prosody measurement lacks precision, reliability, and consistency. Word miscue detection, prominent word detection, and phrasal break detection have precision of 70.4%, 73.2%, and 59.2%, respectively.

1.3 Automatic extraction of prosodic features

Automatic extraction of prosodic characteristics from speech involves several approaches. One approach uses automatic speech recogniser (ASR) for extracting the prosodic features (Shriberg et al., 2005). With ASR, it detects syllables taking into account their pitch, duration, and energy features which are then fed to support vector machines (SVMs). Another approach, which is ASR-free, used stylised pitch dynamics to represent pitch contour (Sonmez et al., 1998). Furthermore, another ASR-free study (Fernandez, 2004) extracts and quantifies loudness, f_0, voice quality and rhythm and assigns features on them as a measure of prosody. This study of Fernandez models prosody using Bayesian networks. A hybrid method (Mary and Yegnanarayana, 2008) uses ASR to merge the syllabic patterns and the ASR-free process to extract their

features. Prosody, as claimed by the researchers can be measured based on changes in f_0, energy, and duration. In assessing children's oral reading prosody, durational features are extracted and analysed using template models and trained generalised models (Duong et al., 2011).

1.4 Oral reading fluency assessments in the Philippines

In the Philippines, the Philippine Informal Reading Inventory (Phil-IRI) of the Department of Education was created for the assessment of reading performance in Filipino and English of elementary school pupils (Llego, 2018). Through an oral reading test, it identifies miscues in oral reading, number of words read per minute and through silent reading test, it assesses comprehension. In general however, informal reading inventories are time-consuming and tedious (Rasinski, 2004), so also is Phil-IRI (Aguilar, 2019). While there is an existing Automated Reading Tutor in Filipino (Pascual and Guevara, 2017), which can detect reading miscues, this cannot classify such miscues nor analyse prosodic features.

1.5 Goals

The aim of this paper is to quantify prosodic properties of children's Filipino read speech, to use these features to assess oral reading fluency, and to characterise prosodic features that are present among fluent readers as well as among non-fluent readers.

This paper is divided into the following: Section 2 presents the methodology and framework of extracting prosodic features from children's Filipino oral reading data. Section 3 presents the results from experiments and investigations, Section 4 contains the analyses, and finally, conclusions together with limitations of the study and recommendations for future work are in Section 5.

2 Methodology

2.1 Framework

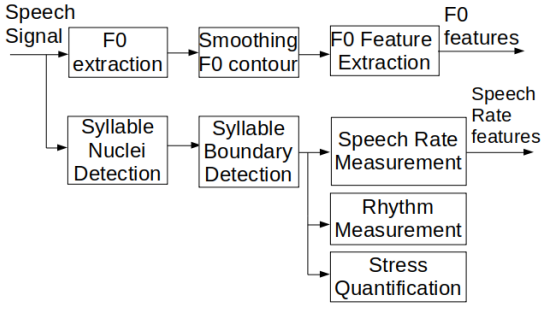
Prosody or prosodic features refer to abstraction of suprasegmentals and their linguistic functions. Pitch and loudness, accentuation, intonation, rhythmic structuring of utterances and speech rate constitute the prosodic attributes of speech. Speech rate can be used to measure fluency (Martins et al., 2007; Ullakonoja, 2009) but that alone is not sufficient (Rasinski, 2004).

In this study, the language of concern is Filipino which is based largely on Tagalog language. So it is logical to consider the properties of Tagalog in order to analyse the prosodic properties of the Filipino language. In terms of speech rhythm, Tagalog is considered syllable-timed (Stockwell, 1957). As to prosody, stress in Tagalog is phonemic. In this case, the primary stress may either be on the last syllable or on the one preceding it. A stress corresponds to vowel duration (lengthening or shortening) (Himmelmann, 2005). Lengthening occurs in a vowel that does not fall at the end of a word. Some words in Tagalog may have the same spelling but have different meanings, usually indicated by differences in accentuation and stress. Proper stress is inferred from

the text being read. As to pitch, objects are usually of higher pitch than the subjects when they are similarly positioned in sentences (Richards, 2017).

This paper focuses on prosody features like speech rate, rhythm and stress patterns, and pitch contours of children's Filipino speech. Figure 1 shows the top level system diagram of extracting prosodic features.

Figure 1 Top level system diagram of extracting prosodic features



It has been established that speech rate as measured by the number of phonemes produced per time unit (Cucchiarini et al., 2002) or as measured by the number of syllables uttered per time unit (Kormos and Dénes, 2004) is a good predictor of subjective reading fluency. Subjective in this sense means that humans are the ones giving oral reading fluency ratings. Along with speech rate itself, some features such as articulation rate and syllable duration were also investigated. In a syllable-timed language, rhythm and stress are exhibited by the variability of vocalic durations, number of pauses and their durations, intensity, and loudness (Fuchs, 2016). With that, rhythm and stress features were extracted within the syllable detection algorithm. This is because the syllable detection method as shown in Figure 3 is basically intensity-based procedure. Furthermore, it can also detect pauses and their durations. The inclusion of pitch features was to find out whether this prosodic feature will be useful and applicable to assessing oral reading fluency in the Filipino language.

2.2 Speech corpus

This study utilised the children Filipino speech corpora similarly used by Pascual and Guevara (2012). It consisted of speech signals recorded from 22 Filipino children reading Filipino texts. The read text passages consisted of 8 grade-level appropriate stories taken from textbooks and children's' books. The audio files were in WAV format, single channel, 16,000 Hz sample rate, 16-bit per sample. There are 75 speech signal files with the total duration of 3 hours, 20 minutes, and 32 seconds.

Speech segments were classified according to story type and were rated by a human rater according to a common 4-point multidimensional fluency scale (Zutell and Rasinski, 1991), see Figure 2. The human rater is an elementary school teacher teaching Reading subjects in Filipino. Furthermore, the human rater also examined the text passages and determined the appropriate pauses during reading. Fluent readers make appropriate pauses at grammatical boundaries while non-fluent readers often do not. The number of determined appropriate pauses was compared with the number of pauses detected in the corresponding speech segments.

Figure 2 Four-point multidimensional oral reading fluency scale

Use the following scales to rate reader fluency on the dimensions of expression and volume, phrasing, smoothness, and pace. Scores range from 4 to 16. Generally, scores below 8 indicate that fluency may be a concern. Scores of 8 or above indicate that the student is making good progress in fluency.

Dimension	1	2	3	4
A. Expression and Volume	Reads with little expression or enthusiasm in voice. Reads words as if simply to get them out. Little sense of trying to make text sound like natural language. Tends to read in a quiet voice.	Some expression. Begins to use voice to make text sound like natural language in some areas of the text, but not others. Focus remains largely on saying the words. Still reads in a quiet voice.	Sounds like natural language throughout the better part of the passage. Occasionally slips into expressionless reading. Voice volume is generally appropriate throughout the text.	Reads with good expression and enthusiasm throughout the text. Sounds like natural language. The reader is able to vary expression and volume to match his/her interpretation of the passage.
B. Phrasing	Monotonic with little sense of phrase boundaries, frequent word-by-word reading.	Frequent two- and three-word phrases giving the impression of choppy reading; improper stress and intonation that fail to mark ends of sentences and clauses.	Mixture of run-ons, mid-sentence pauses for breath, and possibly some choppiness; reasonable stress/intonation.	Generally well phrased, mostly in clause and sentence units, with adequate attention to expression.
C. Smoothness	Frequent extended pauses, hesitations, false starts, sound-outs, repetitions, and/or multiple attempts.	Several "rough spots" in text where extended pauses, hesitations, etc., are more frequent and disruptive.	Occasional breaks in smoothness caused by difficulties with specific words and/or structures.	Generally smooth reading with some breaks, but word and structure difficulties are resolved quickly, usually through self-correction.
D. Pace (during sections of minimal disruption)	Slow and laborious.	Moderately slow.	Uneven mixture of fast and slow reading.	Consistently conversational.

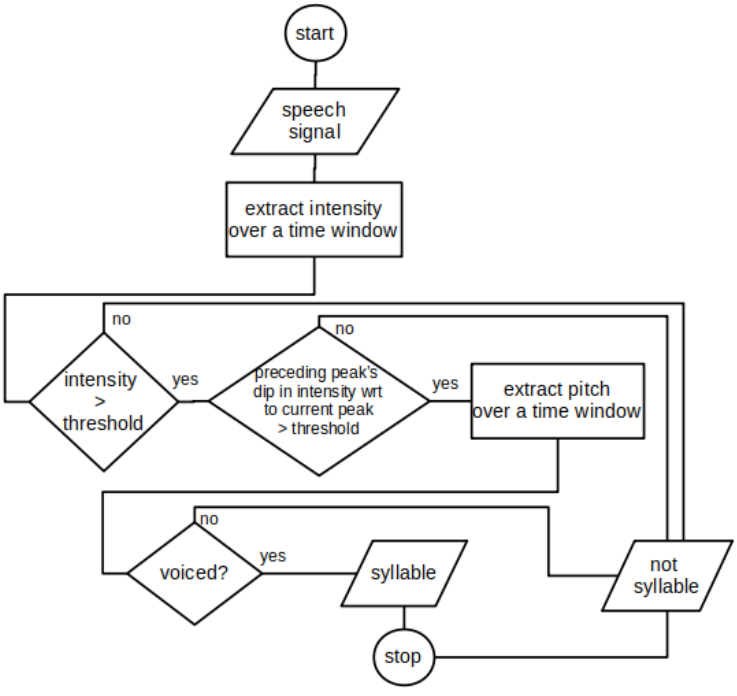
Source: Adapted from "Training Teachers to Attend to Their Students' Oral Reading Fluency," by J. Zutell and T. V. Rasinski, 1991, *Theory Into Practice, 30*, pp. 211-217.

2.3 Syllable nuclei and boundary extraction

Syllable segmentation is done before the process of extracting prosodic features. This involves identifying the syllable nuclei and boundaries. We proceeded by assuming that no hand-labelled annotations and segmental alignments from textual transcriptions are available. In a sense, the algorithms will go blind without the help from automatic speech recogniser.

To extract syllable nuclei and boundaries, Praat software (Boersma andWeenik, 2019), together with a Praat script, modified from Corretge (2019), Quené et al. (2010), and (de Jong and Wempe, 2009), was used on the speech data set. The method uses peaks in intensity to detect the nucleus of syllables. As vowels often get the highest intensity peaks, syllable detection is similar to vowel detection for most cases. For Filipino language, there is a one-to-one relationship between vowels and syllables except for very few words like *ng* (no vowel, pronounced as one syllable) and *mga* (one vowel, pronounced as two syllables). So this syllabication method is appropriate for this language. Figure 3 shows the high-level process of extracting syllables.

Figure 3 Extracting syllables from a speech signal



The output of the script was then interfaced to a MATLAB (2019) script for further processing. Accuracy of syllable detection was determined by comparing the output against hand-counted values. The average of relative errors was then computed. The relative error is given by equation (1).

$$\text{relative error} = \frac{N_C - N_M}{N_M} \times 100\% \tag{1}$$

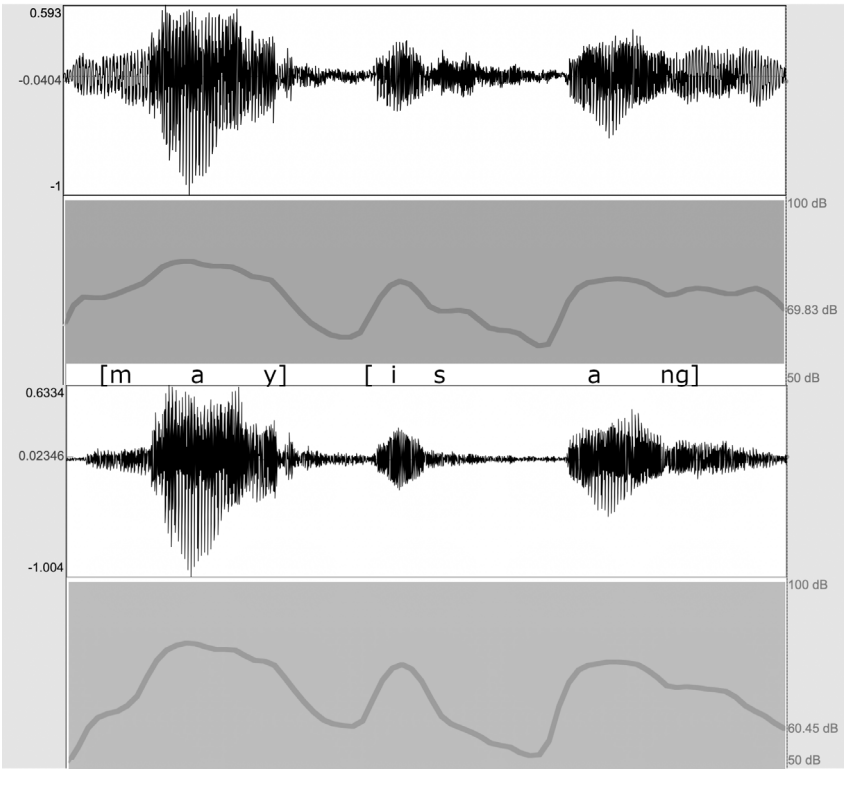
where:

N_C : computed number of syllables

N_M : hand-counted number of syllables.

It was observed that high and low frequency sounds like /s/ and /t/, and /l/, /n/, /ng/, respectively, were falsely detected as syllables. This is due to the fact that these sounds have high energies at high and low frequency bands, respectively. With this, some attenuation at bottom and top frequencies helped. Hence, to further improve the accuracy of syllable detection, a band-pass filter was added as a pre-processing step. This will filter out the energies associated with the fundamental frequencies (F0) and those energies associated with fricatives. As shown in Figure 4, the unfiltered sound "*may isang"* was detected by the script as having 4 syllables while the band-pass filtered one was correctly detected as having 3 syllables. The unfiltered sound has high intensity on the /ng/ sound (the last part of the intensity contour in Figure 4). Hence, the /ng/ sound was detected as a syllable. In contrast, the intensity of the /ng/ sound in the filtered utterance was attenuated.

Figure 4 Waveforms and intensity contours of the utterance "*may isang*". The upper waveform and intensity plot belong to the unfiltered sound while the lower waveform and intensity plot belong to the sound file with band-pass (420–5000 Hz) applied



The parameters in the Praat script were further tweaked in order to determine which values give the minimum average error. Table 1 shows the parameters of the syllable detection script and the range of values for which it was tested. For the lower cutoff frequency of the band-pass filter, the values range from 200 Hz to 450 Hz with increments of 10 Hz. The upper cutoff frequency of the band-pass filter was set to 5000 Hz in order to filter out high frequencies that are not important in speech intelligibility.

Table 1 Parameters of syllable detection script and the ranges for which they were tested for optimisation

Parameter	*Range of values for which optimisation was tested*
Silence threshold (dB)	{–25, –20}
Intensity threshold (dB)	{2.0, 2.1, ⋯, 3.0}
band-pass filter lower limit (Hz)	{200, 210, ⋯, 450}
Width of filter's smoothing region (Hz)	{50,100}

2.4 Speech rate measurements

Having been able to find syllable nuclei and boundaries of a speech signal, the process of which was described in Section 2.3, it would be straightforward to compute speech rate and related metrics. Speech rate is the ratio of the number of syllables uttered with the total utterance duration (equation 2).

$$\text{speech rate} = \frac{\text{no. of syllables}}{\text{duration}} \tag{2}$$

Aside from speech rate, the number of pauses was also recorded as well as the articulation rate. Using Equation 3, articulation rate was computed for each speech file.

$$\text{articulation rate} = \frac{\text{no. of syllables}}{\text{phonation time}} \tag{3}$$

$$\text{phonation time} = \text{time for syllables} = \text{duration} - \text{time for pauses} \tag{4}$$

2.5 Rhythm, stress, and loudness

To detect pauses, a Praat script was used to extract silence boundaries. No distinction was made among detected silences, filled, and breath pauses. The intensity threshold for silences was set at –25 dB and the minimum silence duration was set to 0.30 ms. The number of pauses detected for each speech file was compared with the human-suggested number of pauses. Comparison is done by computing the relative error (equation 5) with respect to the human-suggested number of pauses.

$$\text{relative error} = \frac{N_C - N_H}{N_H} \times 100\% \tag{5}$$

where:

N_C: computed number of pauses

N_H : human-suggested number of pauses.

The automatic syllabication in Section 2.3 gives information about the time location of syllable nuclei and their boundaries. The duration of syllables of a speech segment was extracted together with their mean, average mean deviation, variance, inter-quartile range, normalised pairwise variability index (nPVI) (Grabe and Low, 2008), and coefficient of variation. A Praat script was used to extract the loudness (in *sone* units) of each syllable of a speech segment. The output of the Praat script was inputted to MATLAB for processing the means, mean deviations, variances and inter-quartile ranges. To maximise the audibility and to disregard external factors that affect the loudness of recorded speech segments, data were normalised first by scaling their amplitudes so that their absolute peaks become 0.99.

A sample of mispronounced words from the data set was investigated in terms of their respective syllable durations. Durations were normalised first with respect to the whole duration of articulated words. These durations of mis-stressed syllables were compared to a set of reference articulations considered correct by human rater.

2.6 Pitch

Pitch values were also computed in Praat. These F0 values were then simplified by stylisation to remove irrelevant properties and noise. Stylisation is done by Praat with a frequency resolution of 6 semitones. The stylised pitch values were then processed in MATLAB for the computation of the following features: minimum, maximum, mean deviations from the mean, root-mean-square of deviations from the mean, mean of the high points, mean of low points, inter-quartile range, skewness, kurtosis, number of high points, and number of low points.

2.7 Fluency classification

Kolmogorov-Smirnov test of normality were done on the values of speech rate, articulation rate, pauses relative errors, duration features, loudness features, and pitch features. One-way ANOVA were then employed on quantities that do not differ significantly from that of which is normally distributed; to find differences in values of speech rate, articulation rate, number of pauses, pauses' relative errors, duration features, loudness features, and pitch features among 4 fluency levels. Tukey's HSD tests were further performed in case ANOVA results are significant.

In order to predict fluency levels based on the prosodic features, machine learning classification schemes were employed. Human-rated perceptual fluency rating was used as response variable and the computed prosodic features were used as predictor variables. Classification training and 5-fold cross validation on the data set were done in MATLAB.

Additional pitch features were extracted from a subset of the data set. This subset contains read passages from a story having expressive sentences, interrogatives and exclamations. The speech files were split to per sentence level. MOMEL (Hirst and Espesser, 1993) values were extracted using a Praat script (Hirst, 2007) on these files. The computed MOMEL values were fitted to a curve using spline approximation. Features of the curve such as area under the curve, root mean square and average peak were computed. To account for readers with higher or lower pitch, values were subtracted first by its minimum before computing the area. These feature values were compared to a set of reference values consisting of 4 readers considered as excellent

readers. Both parametric statistical t-tests and non-parametric Mann–Whitney–Wilcoxon rank-sum tests were employed to find the differences among readers versus each value in the set of reference values.

3 Results

3.1 Blind syllabication method

Using the default values of parameters in implementing de Jong's syllabication script yields an average relative error of 12.17%. The optimised syllable detection script was able to obtain an average relative error of 9.81%. The optimum values of parameters are –25 dB, 2.1 dB, 420 Hz, and 100 Hz for silence threshold, intensity threshold, band-pass filter lower limit, and width of filter's smoothing region, respectively.

Figure 5 shows the result of automatic syllabication on a speech sample *Maikli ang kanyang tuka* (Its beak is short). The top row in the figure shows the waveform, the middle row shows the result of automatic syllabication, and the bottom row shows the human-annotated syllabication. Although the number of syllables detected is correct, some syllable boundaries are different from the annotated ones (s1 and *ma*, s2 and *ik*, s5 and *kan*, s6 and *yang*).

Figure 5 Automatic syllabication of a speech sample *Maikli ang kanyang tuka* (Its beak is short)

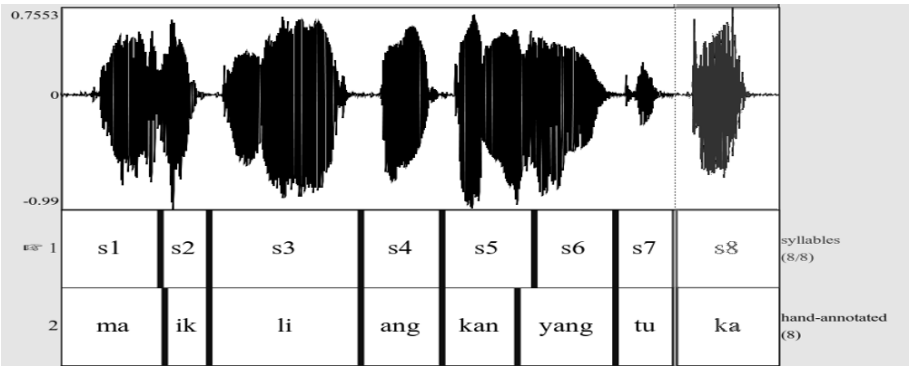
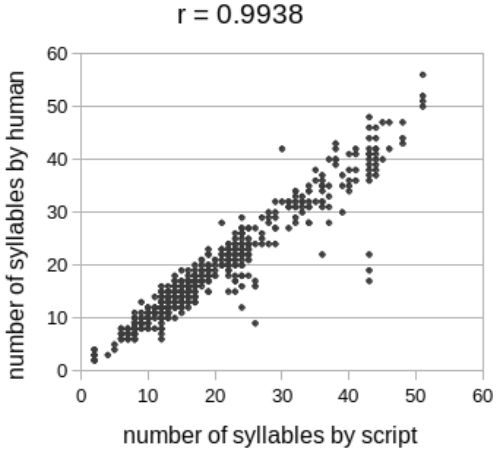


Figure 6 shows the scatter plot of the number of syllables detected by the script versus the number of syllables counted by a human (correlation 0.9938). The number of detected syllables correlates well with the human-counted values.

One of the pre-processing steps in speech signal processing is employing a band-pass filter of 50 Hz to 5000 Hz to reject rumble and high frequency noises. When this pre-processing was employed by the syllabication script on the children Filipino speech corpus, it gave an average relative error of 12%. By inspecting on some speech files where the manual syllable count differs from the script output, it was noticed that some of the false positives belong to some voiced consonants where intensity peaks were considered by the script as syllable. Implementing a band-pass filter between 420 Hz and 5000 Hz gave the minimum average relative error of 9.81%. Also observed was that the occurrences of false positives were in words that contain /ng/ sound. In some cases, the /ng/ is considered a syllable by the script, i.e., with a two-syllable word *lamang* (only),

the script may output three syllables for some speakers. Moreover, when the word has a vowel-vowel combination, e.g. two-syllable word *tao* (man), the script may output only one syllable.

Figure 6 Scatter plot of the number of syllables detected by the script versus number of syllables counted by human



3.2 Speech and articulation rates

Using the Kolmogorov-Smirnov test of normality, values of speech rate, articulation rate, pauses relative errors, duration, loudness, and pitch features mentioned thereafter, do not significantly differ from that which is normally distributed.

Speech rates and articulation rates were observed to be different among groups of fluency levels. Figures 7 and 8 show the box plots of speech and articulation rates among 4 fluency levels, respectively. Speech rate and articulation rate tend to be higher as fluency level becomes better.

Figure 7 Box plot of speech rates among fluency levels

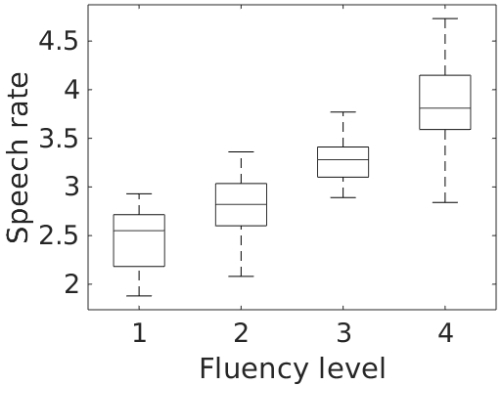
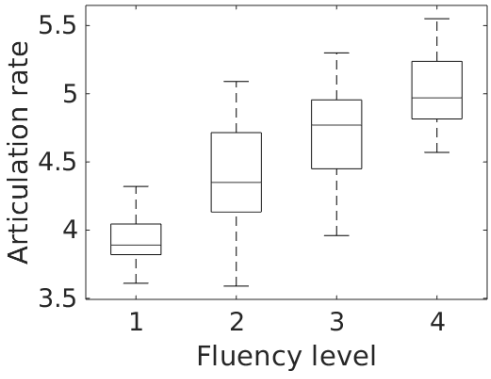


Figure 8 Box plot of articulation rates among fluency levels



At $p = 0.01$ confidence level, analysis of variance suggested that speech rate values differ in one or more fluency levels. The Tukey HSD test further showed that *p*-values in all paired combinations of fluency levels except between level 2 and 1, are all significant, see Table 2. Likewise, for articulation rates, the p-value corresponding to the F-statistic of one-way analysis of variance is also lower than 0.01, suggesting that one or more fluency levels are significantly different. The Tukey HSD test shows that *p*-values in all paired combinations of fluency levels except between level 3 and 2, are all significant.

Table 2 One-way ANOVA and Tukey's HSD test results for speech rate, articulation rate, and relative error of pauses among fluency levels. Column 3 shows the grouping of fluency levels where those that belong to one set are not significantly different

Metric	*p-value*	*Tukey's HSD test*
speech rate	< 0.01	{4}, {3}, {2,1}
articulation rate	< 0.01	{4}, {3,2}, {1}
relative error of pauses	< 0.01	{4}, {3,2}, {1}

Using MATLAB Classification Learner App, features such as speech rate, articulation rate, number of pauses, relative error of pauses, mean, variance, and inter-quartile range of pauses' lengths were considered as predictor variables. It was found that the combination of 3 features, namely, speech rate, articulation rate, and number of pauses, gave the maximum validation accuracy. Also, when the number of fluency levels was reduced to 3 and 2, there was an increased accuracy in the classification. Table 3 shows the 5-fold cross validation accuracy among different number of fluency levels with speech rate, articulation rate, and number of pauses as predictor variables.

Table 3 Classification accuracy of predictor variables: speech rate, articulation rate, and number of pauses among different number of fluency levels

Number of fluency levels	*Classification accuracy*
4	76%
3	85%
2	92%

3.3 Rhythm, stress and loudness

For relative error of pauses, the p-value corresponding to the F-statistic of one-way ANOVA is lower than 0.01 suggesting that relative error of pauses differ in one or more fluency levels. The Tukey HSD test showed that p-values in all paired combinations of fluency levels are all significant, except for level 2 and 3, (see Table 2).

As it was further observed, among fluent readers, the duration of pauses before or after phrase boundaries such as comma, and conjunctions {*at* (and), *pero* (but), *dahil* (because), *kaya't* (and so), *kundi* (but)} was on the average 334 ms. But among frustration readers, the duration, on the average, is 493 ms which is greater than those of fluent readers. The variation of pauses durations among fluent readers is lower (sd = 143 ms) than that of non-fluent readers(sd = 300 ms). Furthermore, normalising the pauses duration with respect to speech rate, fluent readers still have lower values (89 ms versus 217 ms). For intra-sentence pauses durations, that is, between periods, question marks, and interjection marks, fluent readers have an average value of 545 ms while non-fluent readers have 789 ms. The normalised intra-sentence pauses durations have also the same relationship, 144 ms for fluent and 348 ms for non-fluent. The variability of pauses durations is again higher in non-fluent readers (348 ms versus 144 ms). There are instances among non-fluent readers to miss the proper pause durations in both inter- and intra-sentences. The gap between supposed pauses is as low as 10 ms.

The syllable duration metrics tend to be higher in fluency level 1 compared to other levels. Table 4 shows the results of one-way ANOVA for variable mean, average mean deviation, variance, inter-quartile range, normalised pairwise variability index, and coefficient of variation, see Table 4.

Table 4 One-way ANOVA and Tukey's HSD test results for syllables' duration features. Column 3 shows the grouping of fluency levels where those that belong to one set are not significantly different

Metric	*p-value*	*Tukey's HSD test*
mean	< 0.01	{4}, {3}, {2,1}
average mean deviation	< 0.01	{4}, {3}, {2,1}
variance	< 0.01	{4}, {3}, {2,1}
inter-quartile range	< 0.01	{4}, {3}, {2,1}
nPVI	< 0.01	{4, 3, 2}, {1}
coeff. of variation	< 0.01	{4, 3, 2}, {1}

As surveyed in the data set, words that were wrongly stressed are as follows: *abo* (ash), *bangis* (fierce), *bato* (stone), *kalaw* (horn bill), and *tagak* (heron). The word *abo* with correct pronunciation as *abó*, with the stress on the last syllable was pronounced as *ábo*, with the stress on the first syllable, by some non-fluent readers. Similarly, the word *bangís* was mis-stressed as *bángis*, *bató* as *báto*, *kálaw* as *kaláw*, and *tagák* as *tágak*. It was observed that syllable duration is associated with stress; that is, a stressed syllable is usually pronounced longer. By comparing the relative syllable durations within a word, one can detect whether a word is mis-stressed or not. Table 5 shows the ranges of relative percentage durations of syllables within mis-stressed words.

Table 5 Wrongly stressed syllables and their ranges of relative percentage durations within the word

type of readers	*pronunciation*	*range of relative percentage duration of syllables within mis-stressed words*	
non-fluent	*ábo*	*a* 41–66%	*bo* 34–59%
fluent	*abó*	*a* 20–28%	*bo* 72–80%
non-fluent	*bángis*	*ba* 41–43%	*ngis* 57–59%
fluent	*bangís*	*ba* 26–35%	*ngis* 65–74%
non-fluent	*báto*	*ba* 61–84%	*to* 16–39%
fluent	*bató*	*ba* 51–59%	*to* 41–49%
non-fluent	*kaláw*	*ka* 24–28%	*law* 72–76%
fluent	*kálaw*	*ka* 37–49%	*law* 50–63%
non-fluent	*tágak*	*ta* 51–74%	*gak* 26–49%
fluent	*tagák*	*ta* 19–23%	*gak* 77–81%

The metrics for syllable loudness were not significantly different among fluency levels as the *p*-values for syllable loudness' mean, mean deviation, variance, and inter-quartile range were not less than 0.05 in the analysis of variance. The same results were obtained even when the sound segments are not normalised.

The classification training using duration and loudness features yielded less than 60% validation accuracy, except for duration features namely, variance, nPVI, and coefficient of variation. The combination of these 3 features gave the maximum validation accuracy greater than 60%. Table 6 shows the classification accuracy of predictor variables: syllable duration variance, nPVI, and coefficient of variation among different number of fluency levels.

Table 6 Classification accuracy of predictor variables: syllable duration variance, nPVI, and coefficient of variation among different number of fluency levels

Number of fluency levels	*Classification accuracy*
4	74%
3	83%
2	88%

These duration and loudness features such as mean, variance, inter-quartile range, and average mean deviations are considered global in a sense that only one value is given for each speech file. To look into the localised characterisation of syllable durations and loudness, long short-term memory (LSTM) deep learning scheme was employed using MATLAB. Each file was considered as a set of sequence inputs with features namely,

syllable duration, syllable loudness, first- and second-order differences of syllable duration and loudness. Table 7 shows the results of LSTM training with 5-fold cross validation.

Table 7 LSTM classification accuracy of syllable duration, loudness, their delta and delta-delta as predictor features with fluency level as response variable

Number of fluency levels	*Accuracy (%)*					
	Duration features		*Loudness features*		*Combined duration and loudness features*	
	Mean	*Max*	*Mean*	*Max*	*Mean*	*Max*
4	44	60	44	60	50	67
3	46	66	53	66	53	73
2	70	80	68	80	71	80

3.4 Pitch

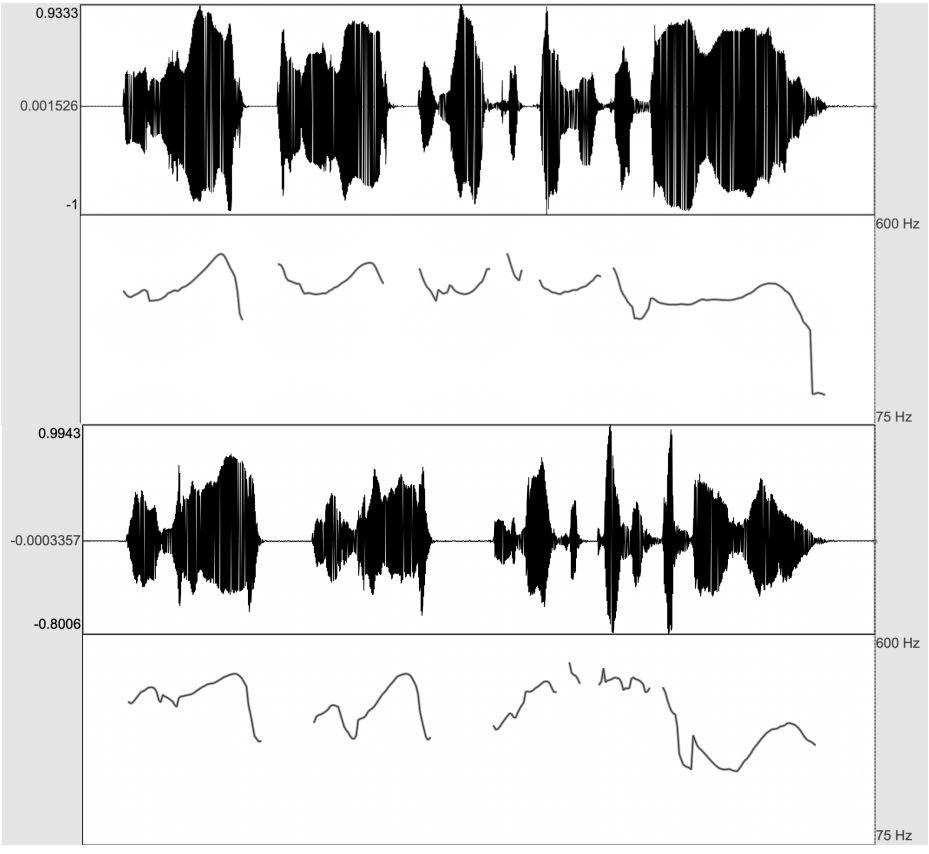
For pitch metrics, a handful of features were computed such as minimum and maximum pitch, average deviations from the mean, average deviations from root-mean-square value, average peaks, average valleys, inter-quartile range, skewness, kurtosis, number of peaks, number of valleys, average of rising slopes between succeeding points, and average falling slopes between succeeding points. Features that were significant were average peak, average valley, skewness, kurtosis, and falling slope average. These statistically significant features were all able to classify only 2 levels of fluency. Table 8 shows the summary of ANOVA results for pitch features that are significant. Average pitch peak values were able to classify most fluent readers from the rest while skewness, kurtosis, and falling slopes were able to classify least fluent readers from the rest. We realised here that pitch can only partially classify oral reading fluency. Those most fluent readers have higher average pitch peaks and those least fluent readers have higher skewness and kurtosis values. Moreover, least fluent readers have flatter average falling slopes.

Table 8 One-way ANOVA for pitch features that are significant

Metric	*p-value*	*Tukey's HSD test*
average peak	< 0.05	{4}, {3, 2, 1}
average valley	< 0.05	{3}, {4, 2, 1}
skewness	< 0.05	{4, 3, 2}, {1}
kurtosis	< 0.05	{4, 3, 2}, {1}
falling slope average	< 0.05	{4, 3, 2}, {1}

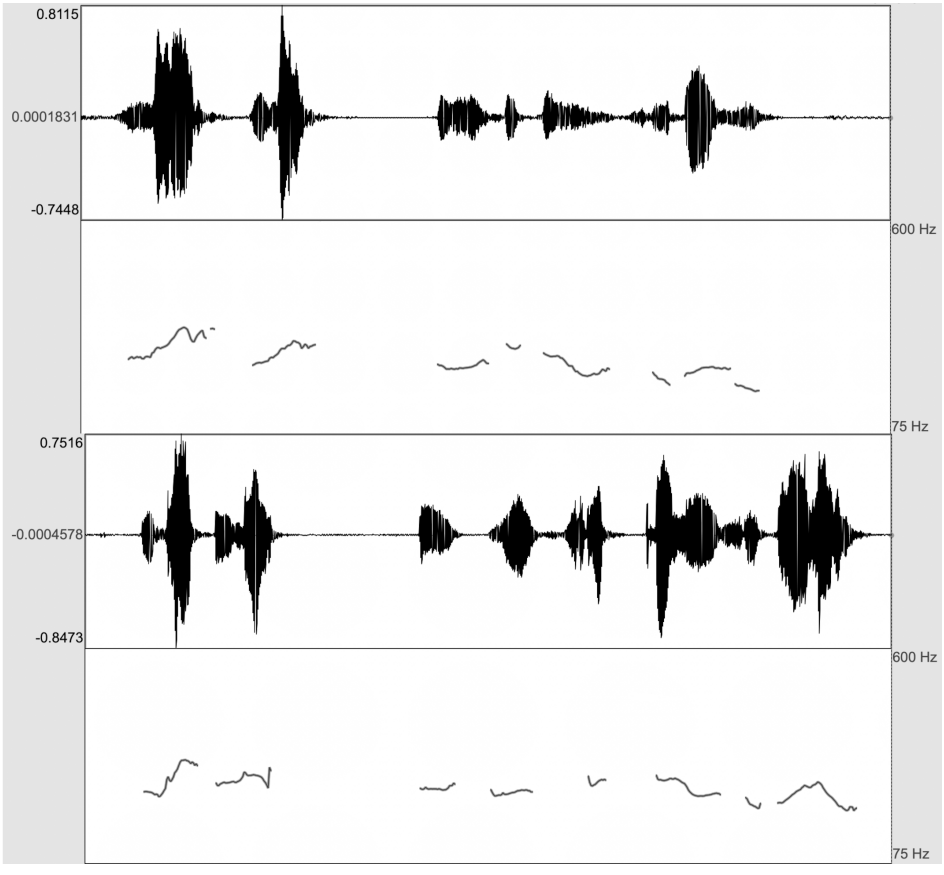
It can be seen that a good reader has a curvy pitch contour profile compared to a non-fluent reader. Figures 9 and 10 show the waveforms and pitch contours of 2 fluent readers versus that of 2 non-fluent readers. Fluent readers have wavy pitch contours that would drop or rise at the end of the phrase boundaries. Non-fluent readers have relatively flat pitch contours compared to fluent readers.

Figure 9 Waveforms and pitch contours of 2 fluent readers reading a sentence, "*Pilo! Pilo! Pumasok ka na sa bahay* (Pilo! Pilo! Please enter the house)



On pitch features belonging to the data set for reading an expressive story, area under the spline curve, root-mean-square and average peak values were able to differentiate non-fluent readers to good readers but only those readers with lowest rating in the fluency rubric. Readers with higher ratings were not statistically different (both using t-test and Mann–Whitney–Wilcoxon rank-sum test) from the good readers data set. Thus, area, root-mean-square and average peak as pitch parameters cannot do a fine-grained classification between fluent and non-fluent readers, at least when using Rasinski's fluency rubric. A Tagalog interrogative exhibits a rising or falling intonation at the end of the sentence depending on type of interrogative. Specifically, only a question which is answerable by yes or no has a rising intonation at the end of the sentence while other types don't have a rising intonation (Pascual et al., 2011). Upon inspection in the data set, fluent readers exhibited a consistent pitch pattern in uttering interrogative sentences. There is a common pattern of rise-fall-rise in their utterance of last two syllables for interrogatives that are answerable by yes or no. Meanwhile, non-fluent readers lack the pronounced rising-falling-rising pattern. Also, the patterns of rising and falling were not consistent among non-fluent readers. Some would do a mixed-up of rising or falling intonation on all types of interrogative sentences.

Figure 10 Waveforms and pitch contours of 2 non-fluent readers reading a sentence, "*Pilo! Pilo! Pumasok ka na sa bahay* (Pilo! Pilo! Please enter the house)



4 Discussions

We proposed a method for syllabication for children's Filipino read speech that requires no training data. Pre-processing the data with a band-pass filter improved its performance. This band-pass filter attenuates those high energy non vowel sounds that would otherwise be detected as syllables. We take a note that a different approach to blind syllabication for the Filipino language might improve detection accuracy. Aside from considering just energy dips and voicedness, as the existing method does, looking at formant energy in glottal closure regions (Krishna et al., 2014) might be useful.

As shown by one-way analyses of variance together with post-hoc Tukey HSD tests, speech rates, articulation rates and number of pauses can be used as predictors of oral reading fluency. Post-hoc test showed that these metrics are significant only for three fluency levels. This number of levels is consistent with what Phil-IRI adapted. Phil-IRI classifies readers as independent, instructional, or frustration. Further results from classification training showed that at 2 and 3 levels of fluency, classification performance has 92% and 85% accuracy, respectively. Hence, when one wants to assess reading

fluency into just two categories (fluent, non-fluent), speech rate, articulation rate, and number of pauses can be used to yield reasonable prediction results.

The tendency of syllable duration metrics to be higher in lower fluency levels can be attributed to disfluencies like lengthenings and hesitations which may imply more effort, tension, or struggle. Thus, low decoding rate or automaticity.

Because pitch variations are innately slight in Tagalog sentences, it is difficult to attribute pitch variations to fluency. Sentences usually begin with normal pitch which gradually rises on stressed syllables and may even rise further in some interrogative sentences. It may end in a normal or slightly lower level if the sentence is declarative (Ramos, 1971).

It is also worth noting that Tagalog is classified as a syllable-timed language (Stockwell, 1957) that is, syllables take approximately equal amounts of time to pronounce. It is also a non-tonal language. Being tonal means having a word to convey a different meaning by changing the tone (pitch) even if the pronunciation of the word is otherwise the same. Tagalog speakers put emphasis by lengthening some syllables. Experiments on syllable durations (Table 4) showed that syllable duration features can classify readers' fluency. Therefore, it is syllable duration, not pitch, that contributes more to prosodic emphasis in Tagalog read speech.

5 Conclusions

In summary, we found out that speech rate, articulation rate, pauses, syllable duration, and to some extent, pitch, as measures of prosody, can be used to predict children's Filipino oral reading fluency. While these metrics were already found to be usual predictors of fluency, (e.g., Morris et al., 2018; Schwanenflugel et al., 2004), we automated the process of syllabication and extraction of prosodic features without the aid of an automatic speech recogniser. We also characterised prosodic features that are useful to the context of children's Filipino read speech. By comparing syllable durations with a set of reference readers, we were able to detect mispronounced words. Although the syllabication script is far from perfect as it has 9.81% average relative error, this benchmark will serve as a baseline for future attempts on automated syllabication in the Filipino language. In particular, the /ng/ sound and vowel-vowel word formations were found to be problematic in syllable detection.

5.1 Limitations

We used a single human rater to perceptually assess oral reading fluency using a commonly used assessment rubric. For a more fine-grained and reliable perceptual assessments, it is recommended to have several human raters. A detailed study should also be done on stressed syllables in combination with the glottal stop. This feature was not investigated here because the data set do not contain miscues involving mispronunciations related to glottal stop. It is our view that the results of this study will serve as contributions to efforts in automating oral reading fluency assessments in the Philippines, thereby helping address reading-related issues that beset basic education in the country.

Acknowledgements

The authors would like to thank the University of the Philippines Digital Signal Processing Laboratory for providing the Children Filipino Speech Corpus as well as the oral reading assessment audio files needed for the experiments. Francis D. Dimzon would like to thank the University of the Philippines Visayas for the fellowship grant support in order to pursue his PhD studies at De La Salle University.

References

Aguilar, Y. (2019) 'On challenges besetting the implementation of Phil-IRI', Personal interview as an elementary school teacher, November.

Alwan, A., Bai, Y., Black, M., Casey, L., Gerosa, M., Heritage, M. et al. (2007) 'A system for technology based assessment of language and literacy in young children: the role of multiple information sources', *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pp.26–30.

Bicol Standard (2019) *Deped rd sadsad to focus on non-readers problem*.

Boersma, P. and Weenik, D. (2019) *Praat (version 6.1*.06) [software]. Available online at: www.praat.org

Bolaños, D., Cole, R., Ward, W., Borts, E. and Svirsky, E. (2011) 'Flora: Fluent oral reading assessment of children's speech', *TSLP*, Vol. 7, p.16.

Corretge, R. (2019) *Praat vocal toolkit: A praat plugin with automated scripts for voice processing*.

Cucchiarini, C., Strik, H. and Boves, L. (2002) 'Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech', *The Journal of the Acoustical Society of America*, Vol. 111, pp.2862–2873.

de Jong, N.H. and Wempe, T. (2009) 'Praat script to detect syllable nuclei and measure speech rate automatically', *Behavior Research Methods*, Vol. 41, No. 2, pp.385–390.

Duong, M., Mostow, J. and Sitaram, S. (2011) 'Two methods for assessing oral reading prosody', *ACM Trans. Speech Lang. Process.*, Vol. 7, No. 4, pp.1–14.

Fernandez, R. (2004) *A Computational Model for the Automatic Recognition of Affect in Speech*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Fuchs, L., Fuchs, D., Hosp, M.K. and Jenkins, J.R. (2001) 'Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis', *Scientific Studies of Reading*, Vol. 5, No. 3, pp.239–256.

Fuchs, R. (2016) *The Concept and Measurement of Speech Rhythm*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp.35–86.

Godde, E., Bailly, G., Escudero, D., Bosse, M. and Gillet-Perret, E. (2017) 'Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings', *WOCCI 2017 - 6th Workshop on Child Computer Interaction*, Glasgow, UK, pp.23–27.

Grabe, E. and Low, E.L. (2008) 'Durational variability in speech and the rhythm class hypothesis', in Gussenhoven, C. and Warner, N. (Eds): *Laboratory Phonology 7*, pp.515–546.

Himmelmann, N. (2005) 'Tagalog', in Adelaar, A. and Himmelmann, N. (Eds): *The Austronesian languages of Asia and Madagascar*, Routledge, London, pp.350–376.

Hirst, D. (2007) 'A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation', *Proceedings of the 16th International Congress of Phonetic Sciences*.

Hirst, D. and Espesser, R. (1993) 'Automatic modelling of fundamental frequency using a quadractic spline function', *Travaux de l'Institut de Phonétique d'Aix*, Vol. 15, pp.75–85.

Kormos, J. and Dénes, M. (2004) 'Exploring measures and perceptions of fluency in the speech of second language learners', *System*, Vol. 32, pp.145–164.

Krishna, H., Mounika, K.V. and Vuppala, A. (2014) 'Improved syllable nuclei detection using formant energy in glottal closure regions', *International Conference on Devices, Circuits and Communications, ICDCCom 2014 - Proceedings*.

Larsen, I. (2016) *Increasing reading fluency in Spanish by doing repeated readings in English*, Master's thesis, California State University, Stanislaus, California.

Llego, M.A. (2018) *Updated Phil-IRI Manual*.

Mary, L. and Yegnanarayana, B. (2008) 'Extraction and representation of prosodic features for language and speaker recognition', *Speech Communication*, Vol. 50, No. 10, pp.782–796.

MATLAB (2019) *Version 9.7.0.1216025 (R2019b) Update 1*, The MathWorks Inc., Natick, Massachusetts.

OECD (2019) *PISA 2018 results: Combined executive summaries*, Vols. 1/2/3.

Pascual, R.M. and Guevara, R.C.L. (2012) 'Developing a children's Filipino speech corpus for application in automatic detection of reading miscues and disfluencies', *TENCON 2012 IEEE Region 10 Conference*, pp.1–6.

Pascual, R.M. and Guevara, R.C.L. (2017) 'Experiments and pilot study evaluating the performance of reading miscue detector and automated reading tutor for filipino: a children's speech technology for improving literacy', *Science Diliman*, Vol. 29, No. 1, pp.5–36.

Pascual, R.M., Rosero, M.W., Nolasco, R. and Guevara, R. (2011) 'Characterization of interrogative sentences in Filipino speech', *Proceedings of the 8th National Natural Language Processing Research Symposium*, De La Salle University, Manila, pp.46–51.

Quené, H., Persoon, I. and de Jong, N. (2016) *Speech rate: Praat script that detects syllable nuclei*.

Ramos, T. (1971) *Tagalog Structures*, University of Hawaii Press, Honolulu, HI, USA.

Rasinski, T. (2004) *Assessing reading fluency*, Honolulu, HI.

Richards, N. (2017) 'Some notes on Tagalog prosody and scrambling', *Glossa: A Journal of General Linguistics*, Vol. 2, p.21.

Sabu, K. and Rao, P. (2018) 'Automatic assessment of children's oral reading using speech recognition and prosody modeling', *CSI Transactions on ICT*, Vol. 6.

SEA-PLM (2019) *Main Regional Report Summary*.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A. and Stolcke, A. (2005) 'Modeling prosodic feature sequences for speaker recognition', *Speech Communication*, Vol. 46, pp.455–472.

Sonmez, M., Shriberg, E., Heck, L. and Weintraub, M. (1998) 'Modeling dynamic prosodic variation for speaker variation', *International Conference on Spoken Language Processing*, Vol. 7, pp.3189–3192.

SRI International (2019) *Eduspeak*.

Stockwell, R. (1957) *A Contrastive Analysis of English and Tagalog*, Vols. 1/2 of *A Contrastive Analysis of English and Tagalog*, Department of English, University of California, Los Angeles, Westwood, California.

Zutell, J. and Rasinski, T.V. (1991) 'Training teachers to attend to their students' oral reading fluency', *Theory Into Practice*, Vol. 30, No. 3, pp.211–217.