SCIENCE DILIMAN (JANUARY-JUNE 2017) 29:1, 5-36

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

Ronald M. Pascual*
Far Eastern University Manila

Rowena Cristina L. Guevara
University of the Philippines Diliman

ABSTRACT

The latest advances in speech processing technology have allowed the development of automated reading tutors (ART) for improving children's literacy. An ART is a computer-assisted learning system based on oral reading fluency (ORF) instruction and automated speech recognition (ASR) technology. However, the design of an ART system is language-specific, and thus, requires developing a system specifically for the Filipino language. In a previous work, the authors have presented the development of the children's Filipino speech corpus (CFSC) for the purpose of designing an ART in Filipino. In this paper, the authors present the evaluation of the ART in Filipino which integrates a reference verification (RV)- and word duration analysis-based reading miscue detector (RMD), a user interface, and a feedback and instruction set. The authors also present the performance evaluation of the RMD in offline tests, and the effectiveness of the ART as shown by the results of the intervention program, a month-long pilot study that involved the use of the ART by a small group of students. Offline test results show that the RMD's performance (i.e., FA rate ≈ 3% and MDerr rate ≈ 5%) is at par with those from state-of-the-art RMDs reported in the literature. The results of the ART intervention experiment showed that the students, on the average, have improved in their words correct per minute (WCPM) rate by 4.66 times, in their ORF-16 scores by 6.0 times, and in their reading comprehension exam scores by 4.4 times, after using the ART.

Key words: Reading miscue detector, automated reading tutor, reference verification, word duration analysis, Filipino speech

*Corresponding Author

ISSN 0115-7809 Print / ISSN 2012-0818 Online

INTRODUCTION

The latest advances in the speech processing technology, coupled with the current problems that the country's primary education system are facing, such as the poor reading performance of the students and the shortage of teachers, inspired the authors to focus on the development of an automated reading tutor (ART) for improving Filipino children's literacy. An ART is a computer-assisted learning system based on oral reading fluency (ORF) instruction and automated speech recognition (ASR) technology. The main task that an ART performs is the automatic detection of reading miscues or disfluencies in an input speech. Through the reading miscue detector (RMD), the ART is capable of "listening" to the reader and spot reading errors so that it may offer help (e.g., by modeling the correct pronunciation of a text passage) whenever necessary. Figure 1 presents an overview of the ART system and its basic components. It must be noted that, unlike a conventional automatic speech recognizer, the RMD knows in advance the desired or target speech pattern that should be uttered by the learner or user. The objective is to identify possible deviations (miscue, error, or disfluency) from the target speech pattern using a certain method.

The designs of ART and RMD systems however are language-specific. For instance, the Project LISTEN's reading tutor (Mostow et al. 1994), the Colorado Literacy Tutor (Hagen et al. 2003), and the system presented by Black et al. (2011) were all designed for the English language. The RMD systems presented by Liu et al. (2008) and by Duchateau et al. (2006) were designed for the Chinese Mandarin and Dutch

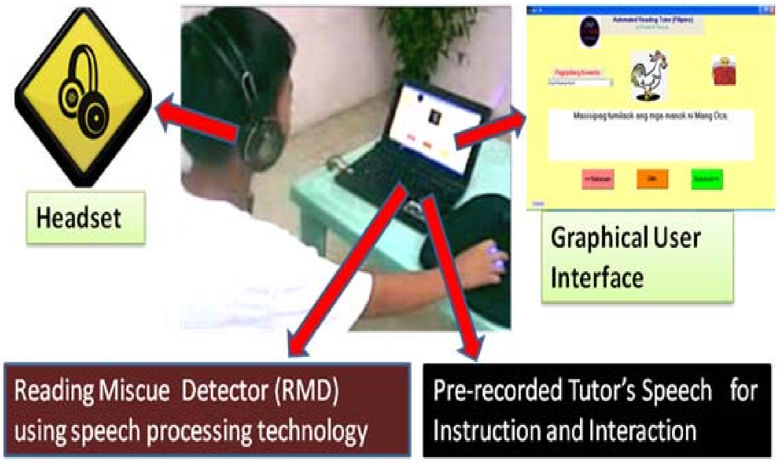

Headset
Graphical User Interface
Reading Miscue Detector (RMD) using speech processing technology
Pre-recorded Tutor's Speech for Instruction and Interaction

Figure 1. Overview of the automated reading tutor (ART) system and its basic components.

languages, respectively. Recently, Rahman et al. (2014) made an effort to develop an ASR system that can be used for ARTs for Malay-speaking children, while Rayner et al. (2014) developed a rule- and grammar-based computer-assisted language learning (CALL) system for German-speaking children.

In this study, the authors focused on the development of a system for Filipino, the national language of the Philippines and a language used in the Philippine basic education system. Features and orthography of Filipino are very distinct from other languages, and thus, there is an apparent need to develop a system specifically designed for the language. For instance, according to speech rhythm, it has been shown that Filipino is generally classified as a syllable-timed language (Guevara et al. 2010). In syllable-timed languages, and unlike in stress-timed languages, such as English, every syllable is perceived as taking up roughly the same amount of time, except for small variations due to the prosody. Moreover, the authors decided to develop a system specifically designed for children in the early grade levels because, according to educators and reading experts, intervention programs, such as reading tutorials done at early grades, are most effective and likely ensure reading success at later grades (Wasik and Slavin 1993; National Reading Panel 2000).

However, a difficulty in developing systems for children is the unavailability of appropriate children's speech corpus that can be used for a particular application in a particular language (Gerosa et al. 2009; Russell 2010). It was noted by Russell (2010) and suggested by Gerosa et al. (2009) that, although an ASR system trained on adults' speech can employ an adaptation technique that improves its performance in processing children's speech, it is unlikely that its performance will exceed that of a counterpart system trained on children's speech. In this study, the authors also present the development of a children's Filipino speech corpus or the CFSC (Pascual and Guevara 2012a) that was used for the design and implementation of the RMD system for Filipino.

Most of the RMD systems that were reported in the literature have generally used either of the two baseline systems: (1) a conventional ASR; or the (2) reference verification (RV) method. The RMD systems presented by Mostow et al. (1994), Liu et al. (2008) and Duchateau et al. (2006) all employed the first type of baseline system (i.e., a conventional ASR) for decoding what the reader has said. In the conventional ASR baseline system type, the recognition results are compared with the "reference" (i.e., the target or expected sound/s in the text to be read) to check whether there are any deviations (i.e., reading miscues or disfluencies). Moreover, a suitable language model (LM) based on the reading text is typically used together with the ASR baseline system, in order to improve the recognition performance.

By contrast, other RMD systems, such as by Black et al. (2011) and by Bolaños et al. (2009), employ the second type of baseline system (i.e., the RV method) and do not use an LM. Under the RV framework, the reader's speech data and the reference are usually forced-aligned while the likelihood of the sounds is calculated. The system then classifies whether an input speech sound, in comparison with the reference, is accepted or rejected. Thus, the RV method may be regarded as a speech classification method rather than a speech recognition method. The RV method may also similarly be seen as a form of speech verification or pronunciation verification. An obvious advantage of the RV method over the conventional ASR baseline system is that it avoids the problems caused by using a complex or dynamic LM (Bolaños et al. 2009). The RV method, however, suffers from relatively lower miscue detection rate due to its observed tendency to usually ignore single phone or syllable deletion, insertion, or substitution, or even repetitions and self-corrections. Thus for practical RMD systems, an additional process is usually integrated to the baseline RV method, in order to achieve a better performance.

In this paper, the authors present the performance evaluation of the RMD system for the automatic detection of reading miscues in children's Filipino speech. The RMD system is the core of the ART system that the authors have developed for Filipino (Pascual and Guevara 2012b). The architecture of the RMD system consists of two levels: (1) a phone-level RV method on the first level; and, (2) a word-level alignment and word duration analysis (WDA) method on the second level. An interesting related study by Duong et al. (2011) discusses about the use of word duration-based template models for automatically assessing children's oral reading prosody in English. The focus of this study is the use of WDA for automatic detection of reading miscues and disfluencies for application in ART system for Filipino.

Over the past few years, a number of field or pilot studies regarding the implementation and evaluation of these ARTs in various languages and countries have been published. For instance, Mills-Tettey et al. (2009) conducted a field study in selected schools in Ghana and Zambia in Africa to investigate the viability and effectiveness of the use of the Project LISTEN's reading tutor, in order to improve the reading skills of children in English as a second language (L2). Using the same ART system, Mostow et al. (2013) made a 7-month study that involved 178 students, and claimed that the use of the ART resulted to improvements expected from guided oral reading, such as higher gains in fluency and reading comprehension. Duchateau et al. (2009) presented the evaluation of an ART for fluency instruction in Dutch as a first language (L1), and claimed that the specific ART works satisfactorily for children, even for those with reading disabilities, in a real school environment. Tsau (2012) conducted a field study, which was participated in by students in central Taiwan, and reported that their ART system named "My English Tutor" have successfully enhanced the oral reading fluency (ORF) of the EFL learners. Similarly, Reeder et al. (2015) employed an ART system for English in their study conducted in a public elementary school in Vancouver, Canada, and concluded that the ART system successfully contributed to the reading development of young learners of English as additional language (EAL).

In this paper, the authors present the evaluation of an ART for Filipino that integrates an RMD, a user interface, and a feedback and instruction set. The next sections present a discussion of the results of the ART intervention program, a month-long pilot study that involved the use of the ART by a small group of students. The program aims to evaluate the effectiveness of the ART in improving the reading skills of the learners.

DESIGN AND EVALUATION METHOD FOR THE READING MISCUE DETECTOR FOR FILIPINO

Children's Filipino Speech Corpus and Models

Some studies in the past few years, such as by Kazemzadeh et al. (2005), Batliner et al. (2005), Cleuren et al. (2008), and Gao et al. (2012), have focused on the development of children's speech corpora in languages, such as English, Dutch, Italian, German, Swedish, and Mandarin. The absence of a speech corpus that can be used for the development of an RMD and ART for Filipino has motivated the authors to develop a medium-scale, gender- and age-balanced CFSC (Pascual and Guevara 2012). Nearly all of the speakers in the CFSC are native speakers of Filipino.

The CFSC consists of two parts: (1) a part containing good reading pronunciations; and, (2) a part containing examples of actual reading miscues and disfluencies. The CFSC provides the following data sets: training data set for the generation of speech models, reference speech features (such as word durations) set extracted from good pronunciations, offline test set for the evaluation of the RMD system, and data set for the analysis of actual reading miscues found in children's Filipino speech.

The first part of the children's speech corpus (i.e., the good pronunciations) contains about five hours of continuous read speech collected from a total of 37 Grades 2 to 5 students (ages ranging from 7 to 12 years). Out of the 37 students, 17 are girls and 20 are boys. The second part of the children's speech corpus (i.e., the part containing reading miscues) contains about three hours of continuous read speech collected from a total of 20 Grades 1 to 3 students (ages ranging from about 6 to 9 years). Of these students, 11 are girls and 9 are boys.

Nearly the entire CFSC contains orthographic transcriptions for the speech data. To further make the CFSC useful for the RMD design, the authors transcribed a part of the CFSC at phoneme level. Note that the smallest possible sound unit where a reading miscue may occur is in a single phone or syllable, thus suggesting the need for phone-level transcriptions. The phone-level transcription process, which is one of the most expensive parts of the design, was executed in a semi-automated method. That is, the speech data were initially machine-transcribed through phoneme forced-alignment method using a hidden Markov model- or HMM-based speech modeling toolkit HTK (Young et al. 2006), and the machine transcriptions were then manually re-aligned afterwards if needed. A bootstrap data set (about 20 minutes of hand-transcribed speech) was used to facilitate the automated transcription method. The phoneme set used for this study includes a total of 35 phones and diphones as listed in Table 1.

Table 1. Phoneme set used for children's Filipino speech corpus (CFSC) transcriptions

Phone Class	Phones / Diphones
Stop	/p/, /b/, /t/, /d/, /k/, /g/, /q/ (glottal stop)
Fricative	/f/, /v/, /s/, /z/, /sh/
Affricate	/j/
Nasal	/m/, /n/, /ng/
Lateral Liquid	/l/
Retroflex Liquid	/r/
Glide	/w/, /y/
Vowels	/a/, /e/, /i/, /o/, /u/
Diphones	/ha/, /he/, /hi/, /ho/, /hu/, /at/, /aw/, /ay/, /oy/
Pause / Silence	/pau/

A total of nine Filipino text passages, six of which are short stories while the other three are expository-type texts, were used for the CFSC recordings. Adarna House, a publisher of Filipino short stories for children in the Philippines, provided the six short stories while the other three expository-type texts were adopted from various school textbooks. All the text passages are age-appropriate and have been suggested by research collaborators from the College of Education at the University of the Philippines Diliman.

Design of the Reference Verification- and Word Duration Analysis-based Reading Miscue Detector for Filipino

For an RMD, the desired or target speech pattern (reference) that should be uttered by the reader is known in advance, and the objective is to identify possible deviation (miscue or error) from the reference. As mentioned in section I, the RV method for detecting reading miscues aligns the input speech with the reference, and decides through a certain similarity measure whether or not the input speech sounds are the same as the reference sounds.

There are generally two possible approaches in estimating the likelihood or similarity of the expected sounds in the reference to those sounds found in the input speech. The first approach is through a time-domain template matching, while the second approach is through speech model (HMM) comparison using parametric representation of speech features, such as the Mel frequency cepstral coefficients (MFCCs). While template matching-based ASR offers simplicity, it also suffers from several difficulties and limitations, such as inability to generalize features from many speakers and example utterances, low computational efficiency for larger number of test/reference patterns, and the inability to incorporate statistical features from a given data set. The advantages of HMM-based ASR over template matching-based systems have influenced us in selecting the HMM-based approach for designing the RV-based RMD in Filipino. In particular, the HMM-based approach allowed us to practically use all the information available in the training data and to design an RMD that is speaker-independent. Furthermore, the HMM-based approach also permitted us to implement a computationally efficient and practically realizable RMD.

In this study, the authors' initial approach is to design a baseline system that uses phone-level RV method. To do this, the authors first generated the hidden Markov models (HMMs) for all the phones in the phoneme set used in this study by training the system with 1.5 hours of phone-level transcribed children's speech in the CFSC. The HMMs consist of three states with 39 MFCCs comprising 13 static coefficients plus 13 delta coefficients plus 13 acceleration coefficients. An overview of the Markov model is illustrated in Figure 2. It is worth noting that the 3-state HMM prototype shown in Figure 2 appears to actually have five states. This is only due to the speech modeling toolkit convention. The first and the last states are non-emitting states, and thus, are part of the network of HMMs, but do not describe any of the input data. Only the middle three states emit observation vectors.

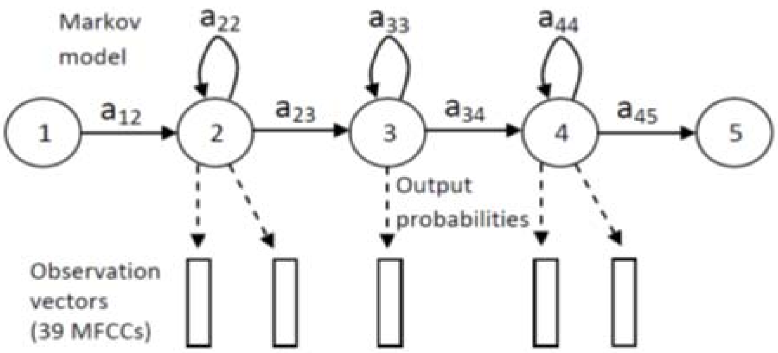


Figure 2. Three-state left-to-right hidden Markov model (HMM) used in this study. States 1 and 5 are non-emitting. Observation vectors consist of 39 Mel frequency cepstral coefficients (MFCCs). The aij's are the transition probabilities.

Given the reference text passage and an input speech data, the phone-level RV method then proceeds as follows:

1. Form the reference (or expected) phone symbol sequence (including symbols for expected short pauses/silences for proper phrasing) from the reference text passage.

2. Perform an HMM Viterbi-forced alignment process between the reference phones and the phones found in the input speech data. (This process produces the log likelihood scores for each phone in the reference).

3. Using the output likelihood scores ρ_v for each phone v from step 2, perform a threshold-based classification to decide whether or not a reading miscue has occurred. That is,

$$md_{RV} = \begin{cases} 1 \quad (miscue\ present), & \min_{all\ v}[\rho_v] < P \\ 0 \quad (miscue\ absent), & \min_{all\ v}[\rho_v] \geq P \end{cases} \qquad (1),$$

where P = log likelihood score threshold.

Figure 3 graphically illustrates the RV method through a specific example. In this example, the reference text is a four-word Filipino phrase /Maraming prutas at gulay/ (Many fruits and vegetables). The RV process is initiated by phonetically spelling out the reference, thus producing the phone sequence: /m/, /a/, /r/, /a/, /m/, /i/, /ng/, /p/, /r/, /u/, /t/, /a/, /s/, /q/, /at/, /g/, /u/, /l/, and, /ay/. The reference phone sequence is then stored as a text file. During the execution of the RV process, the reference phones are all initially assigned to a starting position and have equal durations. The Viterbi-forced alignment process proceeds by taking the reference phones one-by-one and finding the best alignment (i.e., the alignment that produces the maximum likelihood score) with the input speech data. The numbers alongside the phone symbols in the last tier shown in Figure 3 are the log likelihood scores.

Intuitively, we may decide that there has been a reading miscue (i.e., deviation from the expected or reference phones) if a phone likelihood score goes below the lower threshold. While the previous condition generally applies, it is now worth noting that due to the nature of the Viterbi-forced alignment and likelihood scoring, there is also a need to set another threshold (i.e., an upper threshold) for the purpose of detecting a miscue. The upper threshold is necessary for the detection of cases wherein the likelihood scores become too high due to certain types of miscues or disfluencies, such as vowel elongation or pause/silence prolongation.

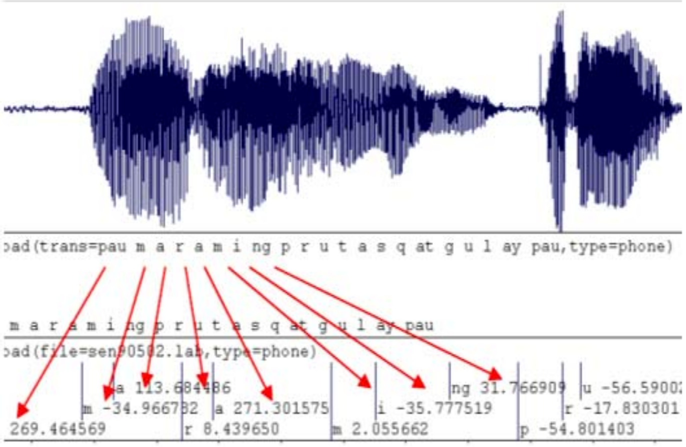

pad(trans=pau m a r a m i ng p r u t a s q at g u l ay pau,type=phone)
m a r a m i ng p r u t a s q at g u l ay pau
pad(file=sen00502.lab,type=phone)
a 113.684486
ng 31.766909
u -56.5900
m -34.966732
a 271.301575
i -35.777519
r -17.830301
269.464569
r 8.439650
m 2.055662
p -54.801403

Figure 3. Illustration of the application of the reference verification (RV) process used in this study to a specific utterance of a sentence in the Filipino reading text.

As mentioned in the previous section, the phone-level RV method alone is expected to give a relatively low miscue detection rate at an acceptable false alarm rate due to its poor ability to detect single-phone or single-syllable errors, as well as disfluencies like brief hesitation pauses, self-corrections, and repetitions. To address this problem, the authors' approach was to integrate a second-level (i.e., word-level) process that uses a duration-based prosodic feature (i.e., word durations) to the baseline phone-level RV process, in order to obtain a better detection of reading miscues. The idea was based on an initial observation that, in many cases of actual reading miscues and disfluencies found in the CFSC, the effective word durations highly deviated from the expected word durations, which are based on good pronunciation examples. Figure 4 illustrates such a case of a reading miscue found in a certain five-word sentence in the CFSC. Figure 4 shows the speech waveforms of a good pronunciation example (upper plot) and an utterance containing a reading miscue (lower plot) for a particular sentence. The machine-detected word boundaries, indicated by the rectangle edges, are shown beneath the respective plots. The orthographic transcriptions of the speech data are also shown below each plot. Note that the duration of word number 3 in the lower plot is significantly longer than that of the upper plot. The observed word duration deviation from that of the good pronunciation is due to the reading miscue (which is a substitution of the word "ito'y" for "itong", followed by a self-correction) found in the lower plot.

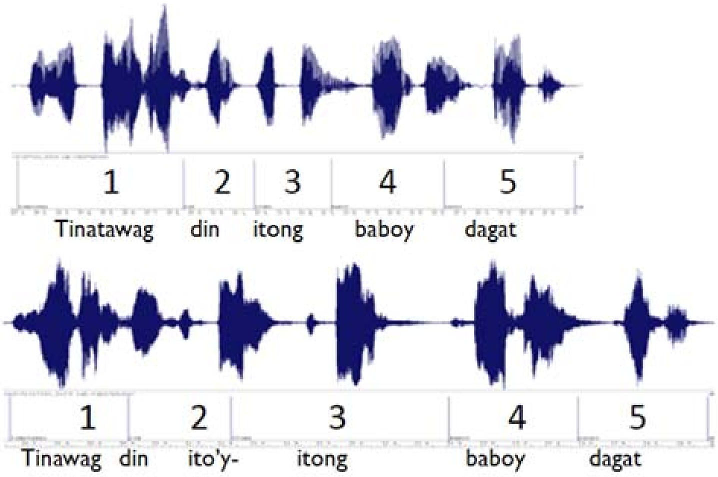


Figure 4. Illustration of the machine-detected word duration deviation (word number 3) from the reference (upper plot) due to a reading miscue found in the speech data shown in the lower plot.

In order to implement the WDA for the purpose of reading miscue detection, the authors initially extracted the average sentence-normalized word durations from the good pronunciation examples in the CFSC. The reference word durations were initially stored in a look-up table. After the initialization process, the main WDA process then proceeds as follows:

1. Given a certain sentence, form the reference word sequence (i.e., a static word decoding network where word-ends are connected only to the fan-in nodes of the alternative pronunciations of the next word appearing in the reference). For the purpose of this step, the authors used a pronunciation dictionary of 650 unique words, including alternative pronunciations and a silence/pause model, found in the reading text passages.

2. Perform a word-level Viterbi-forced alignment process between the reference and the input speech data. (This process produces the machine-detected word boundaries for each word in the reference sentence.)

3. From the machine-detected word boundaries in step 2, calculate the sentence-normalized durations for each word in the sentence. (Normalization with respect to local sentence duration is necessary to compensate for different reading rates.)

4. Calculate the relative deviations (from the normalized reference word durations) of the normalized word durations found in step 3.

5. Using the relative word-duration deviations δ_v for each word *v* in step 4, perform a decision process (of whether or not there was a reading miscue) as follows:

$$md_{WDA} = \begin{cases} 1 \quad (miscue\ present), & \max\limits_{all\ v}[\delta_v] > \Delta \\ 0 \quad (miscue\ absent), & \max\limits_{all\ v}[\delta_v] \le \Delta \end{cases} \qquad (2),$$

where Δ = word duration deviation threshold.

Figure 5 shows the combined phone-level RV and WDA method, which is simply referred to here as the "RV-plus-WDA" method, for the final design of the RMD for Filipino. For the RV-plus-WDA method, the miscue detection is now given by the following classification scheme:

$$md = \begin{cases} 1 \quad (miscue\ present), \quad [md_{RV} + md_{WDA}] > 0 \\ 0 \quad (miscue\ absent), \quad [md_{RV} + md_{WDA}] = 0 \end{cases} \qquad (3),$$

where mdRV and mdWDA are given respectively in Equations (1) and (2).

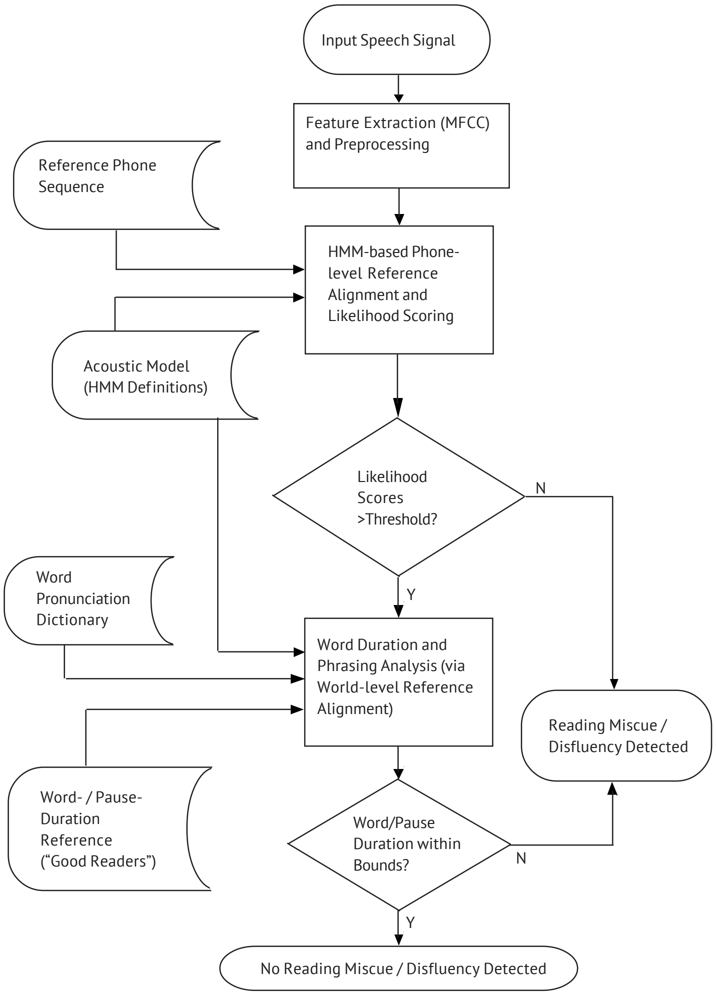

Input Speech Signal
Feature Extraction (MFCC) and Preprocessing
Reference Phone Sequence
HMM-based Phone-level Reference Alignment and Likelihood Scoring
Acoustic Model (HMM Definitions)
Likelihood Scores >Threshold?
N
Y
Word Pronunciation Dictionary
Word Duration and Phrasing Analysis (via World-level Reference Alignment)
Reading Miscue / Disfluency Detected
Word- / Pause-Duration Reference ("Good Readers")
Word/Pause Duration within Bounds?
N
Y
No Reading Miscue / Disfluency Detected

Figure 5. Overview of the combined methods, reference verification (RV) and word duration analysis (WDA), for the reading miscue detector (RMD) design.

Reading Miscue Detector Performance Evaluation Method

In order to evaluate the performances of the RMD systems presented in the previous section, the authors performed two sets of offline tests that employ various threshold values. The first set of tests was performed to evaluate the performance of the phone-level RV-based RMD, while the second sets of tests evaluate the performance of the two-level RV-plus-WDA-based RMD.

For both sets of tests, the authors used an offline test set containing 100 sentences or a total of 1,030 words that were randomly selected from the CFSC. The authors considered selecting the test files, such that there is a fairly balanced representation in terms of gender and of age. Among the 100 sentences in the aforementioned test set, 50 contained at least one reading miscue or disfluency, while the other 50 contain good pronunciations. Analysis of the test set showed that there are seven types of reading miscues found in children's Filipino speech: (1) partial word; (2) hesitation pause; (3) insertion; (4) repetition; (5) substitution; (6) deletion; and, (7) elongation. Table 2 summarizes the relative frequencies of occurrences of the aforementioned reading miscues in the test set.

Table 2. Reading miscue occurrences in the test set

Type of Miscue/Disfluency	Relative Frequency
Hesitation Pause	30.1%
Partial Word	20.4%
Insertion	18.4%
Repetition	13.6%
Substitution	7.8%
Deletion	6.8%
Elongation	2.9%

The performances of RMD systems were evaluated using the two measures commonly used in literature: the false alarm (FA) rate; and, the reading miscue detection error (MDerr) rate (Duchateau et al. 2006; Liu et al. 2008; Black et al. 2011).

The FA rate is defined as the number of words erroneously detected as read incorrectly divided by the total number of correct pronunciations. That is,

$$FA = FP / (TN + FP) \qquad (4),$$

where *FP* = number of false positives (i.e., false detections of a miscue), and *TN* = number of true negatives (i.e., correct detections of the absence of a miscue).

The MDerr rate, also referred to as misdetection (MD) rate, is defined as the number of miscues that were not detected divided by the total number of miscues. That is,

$$MD = FN / (TP + FN) \qquad (5),$$

where *FN* = number of false negatives (i.e., undetected miscues), and *TP* = number of true positives (correctly detected miscues).

READING MISCUE DETECTOR PERFORMANCE: TEST RESULTS AND DISCUSSION

To commence with the first set of offline tests that evaluate the performance of the phone-level RV-based RMD, the authors performed offline tests using the test set described in the previous section for various upper threshold values. For the offline tests, an initial fixed lower threshold value of -1000 (log likelihood score) was employed based on the observation that this setting did not introduce any false alarm. Figure 6 shows the results of the test runs in terms of FA and MDerr rates for various upper threshold values. Figure 6 shows that false alarms vanished at around an upper threshold value of 550. Since the upper threshold generally imposes a less strict condition than that of the lower threshold, the authors decided to employ the aforementioned value as a fixed upper threshold value for all the

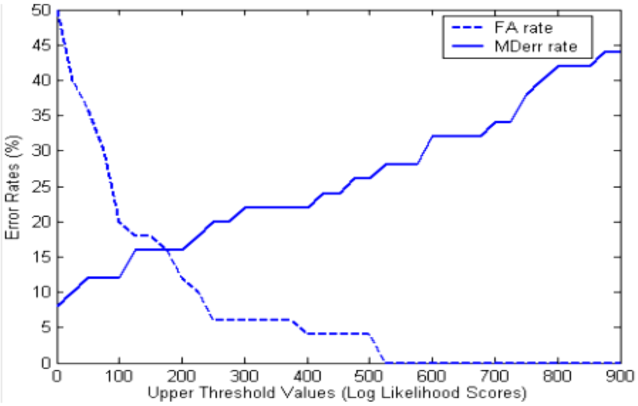


Figure 6. False alarm (FA) and miscue detection error (MDerr) rates as functions of the upper threshold values for the reference verification (RV)-based reading miscue detector (RMD).

succeeding tests and system operations. Thus, for the rest of this article, the term "threshold" for log likelihood score actually pertains to the lower threshold.

In addition, we may note that the plots in Figure 6 contain fluctuations due to the finite amount of data in the offline test set. For the error rate curves in the succeeding figures, the authors minimized the fluctuations, in order to predict the general behavior of the system as the size of the test set increases. That is, the FA and MDerr rate curves were modeled based on the widely used assumption in literature that the probability distribution function (PDF) of the reading miscue characteristics in the test data follows a normal (Gaussian) distribution. Thus, the authors fit the error rate curves to an approximation of a Gaussian cumulative distribution function (CDF) in the least-squares sense.

After setting the upper threshold value for the RMD constant, the authors performed another set of offline tests for various lower threshold values. Figure 7 shows the resulting performance of the of phone-level RV-based RMD in terms of FA (dashed curve) and MDerr (solid curve) rates for various phone likelihood score threshold (i.e., lower threshold) values. The trends in Figure 7 show a generally increasing FA rate and decreasing MDerr rate as the threshold is increased. Since RMDs are never perfect, it has been customary for ART systems to be biased towards having lower FA rates at the expense of having higher MDerr rates. This is done, in order to avoid frustration on the reader with too many unnecessary interventions (Mostow and Aist 1999). Typically, an FA rate equal to or higher than 10% for reading tutors is not a good performance compared to state-of-the-art systems that usually have

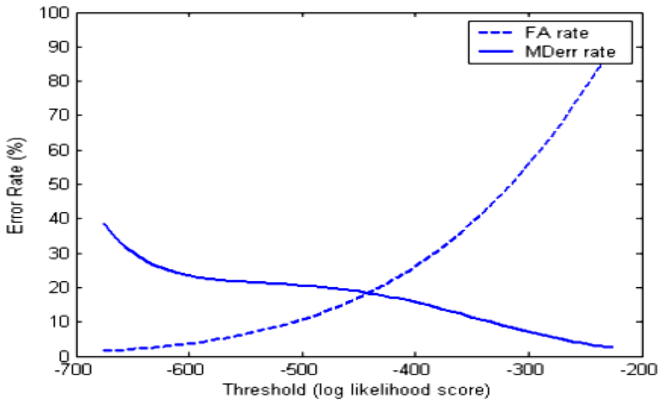


Figure 7. False alarm (FA) and miscue detection error (MDerr) rates of the phone-level reference verification (RV)-based reading miscue detector (RMD).

lower FA rates. Taking into consideration Figure 7, for instance, the probable best case is to adjust the threshold, such that the FA rate is approximately 5%, while the MDerr rate is approximately 22.5%. However, an MDerr rate of 22.5% may generally be seen as still an unsatisfactory performance by an RMD. Thus, the succeeding parts of this section present how much improvement for the RMD's performance can be achieved by incorporating a second level process for reading miscue detection.

Investigation of the misdetection cases revealed that the previously presented baseline method is generally unable to detect the following miscues and disfluencies: (1) single-syllable or single-phone deletion, insertion, or substitution that has durations of 150-250 milliseconds; (2) brief hesitation pause that is less than 500 milliseconds; and, (3) some restarts or self-corrections.

The second set of offline test results shown in Figure 8 presents the performance of the two-level RV-plus-WDA-based RMD in terms of FA (dashed) and MDerr (solid) rates for various word duration deviation threshold values. Note that, unlike with those from Figure 7, the trends in Figure 8 show a generally decreasing FA rate and increasing MDerr rate as the word duration deviation threshold is increased. Note that higher word duration deviation threshold values result to a less strict miscue detector, while the opposite is true for higher likelihood score threshold values.

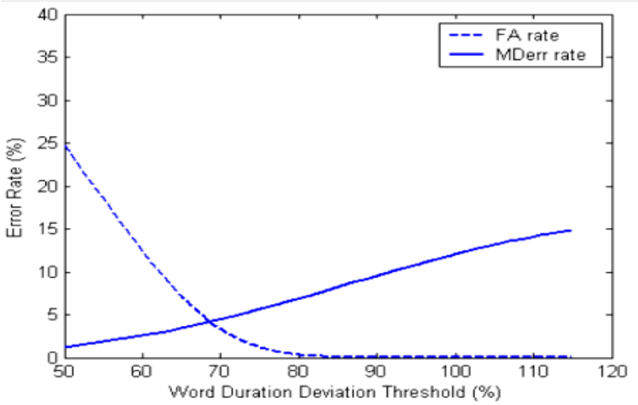


Figure 8. False alarm (FA) and miscue detection error (MDerr) rates for the two-level reference verification and word duration analysis (RV-plus-WDA)-based reading miscue detector (RMD).

As discussed in the previous section, the two-level RV-plus-WDA method combines two different methods that use two different thresholds. The results shown in Figure 8 imply that the word duration deviation threshold varied while the likelihood score threshold remained fixed. The authors have attempted the use of different combinations of the two thresholds. The results of the aforementioned experiments showed that the best results (i.e., lowest overall FA and MDerr rates) were obtained by making the first detector level (i.e., the phone-level RV method) less strict while allowing the second detector level (i.e., the WDA method) catch the misdetection cases in the former. Specifically, the authors have fixed the phone likelihood score threshold, such that the phone-level RV method alone has an FA rate of about 2% at an MDerr rate of roughly 30%.

The final threshold values used for the first and for the second RMD levels, respectively, are: P = -650 (log likelihood score), and Δ = 70% (word duration deviation). This combination seems to give the lowest overall FA and MDerr rates while having a good FA-to-MDerr rate ratio. In particular, the approximate error rates for the threshold combination are FA rate = 3% and MDerr rate = 5%. The FA and MDerr rates for the selected threshold combination may also be deduced from the tabular summary of the various combinations of the threshold values. Table 3 may also be used as a guide in predicting how the RMD system will perform in case another threshold combination is desired.

Table 3. False alarm (FA) and miscue detection error (MDerr) rates for various combinations of two thresholds, P and Δ

	Δ = 50%		**Δ = 755%**		**Δ = 100%**	
	FA rate	**MDerr rate**	**FA rate**	**MDerr rate**	**FA rate**	**MDerr rate**
P = - 1000	20%	4%	2%	8%	0%	12%
P = - 650	20%	2%	2%	6%	0%	10%
P = - 570	20%	0%	6%	2%	4%	8%
P = - 500	26%	0%	12%	0%	10%	6%

Note. P = Phone-Level Log Likelihood Score Threshold; Δ **=** Word Duration Deviation Threshold

In order to obtain a better comparison (independent of threshold) of the system performances, the receiver operating characteristic (ROC) graphs, as shown in Figure 9, were generated by plotting the FA rates versus the MDerr rates. Figure 9 shows the ROC graphs of the phone-level RV-based RMD (dashed curve) and the two-level RV-plus-WDA-based RMD (solid curve). Compared with the phone-level RV method alone, the combined RV-plus-WDA method has significantly improved the RMD's performance. Specifically, Figure 9 shows that, at an FA rate of about 3%, the RV method alone obtained an MDerr rate of about 25%, while the RV-plus-WDA method obtained an MDerr rate of about 5%. Thus, at this FA rate, the combined RV-plus-WDA method provided an MDerr rate absolute improvement of 20% over the RV method alone.

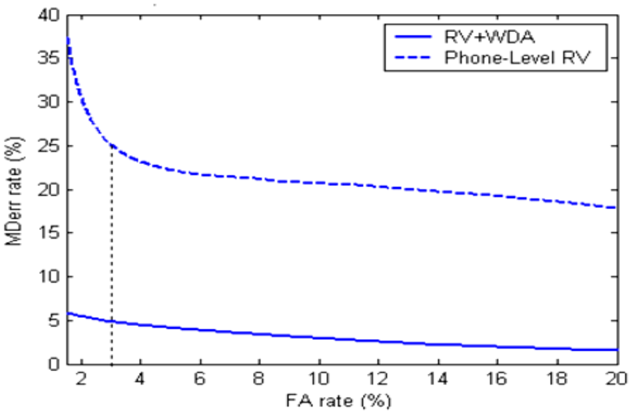


Figure 9. Receiver operating characteristic (ROC) graphs or error rates trade-off curves for the phone-level reference verification (RV) method (dashed), and for the the two-level reference verification and word duration analysis (RV-plus-WDA) method (solid) for the reading miscue detector (RMD) design.

Table 4 summarizes and compares the performances of the phone-level RV and the RV-plus-WDA method at selected operating points. Operating points 1 and 2 are where both methods obtained the same FA rates, namely 3% and 10%, respectively. Operating point 3 is known as the equal error rate (EER), where FA and MDerr rates are equal for a certain method.

Table 4. False alarm (FA) and miscue detection error (MDerr) rates for the phone-level reference verification (RV), and for the two-level reference verification and word duration analysis (RV-plus-WDA) methods at selected operating points

RMD Method	Error Rate	Operating Point 1	Operating Point 2	Operating Point 3
Phone-level RV	FA	3%	10%	18%
	MDerr	25%	20.5%	18%
RV-plus-WDA	FA	3%	10%	4.25%
	MDerr	5%	3%	4.25%

The previous discussions have made clear that the RV-plus-WDA method has a more superior performance than the phone-level RV method. Nearly about 70% of the reading miscues that were missed by the phone-level RV method were successfully detected by the RV-plus-WDA method. One reason that could explain this result is the inability of the phone-level RV method to detect deviations from the expected sounds when the deviations happen only for a short period of time. In particular, the phone-level RV method was observed to have difficulties in detecting syllable insertions, deletions, or substitutions, and word restarts or immediate self-corrections. Upon further investigation, the authors found out that about 57% of the misdetection cases are syllable or phone insertions, substitutions, and deletions. Moreover, the durations of the undetected insertions and substitutions mostly range from about 150 to 250 milliseconds. Brief hesitation pauses, ranging from about 250 to 400 milliseconds, constitute about 28% of the misdetection cases. Self-corrections, restarts, and repetitions all together make up about 14% of the misdetection cases. With these reading miscues, the behavior of the phone alignment method is to align a reference phone to within a beam of few phones in sequence found in the input speech. In the process of seeking alignment, the system has the tendency to either skip some inserted phones or ignore missing phones in the input speech.

Figure 10 shows an example of a reading miscue that was undetected by the phone-level RV method. As we can see from the first plot in Figure 10, the reference phones that were forced-aligned with the input speech are for the Filipino phrase /Tinatawag din itong/. The transcription of the actual input speech shown in the first plot however is given as /Tina(ta)wag din (itoy-) itong/, which contains a syllable deletion (i.e., syllable /ta/ in /Tinatawag/) and a self-correction for a miscue (i.e., /itoy/). Examination of the log likelihood scores, shown at the bottom of the first plot, reveals that the miscues were undetected (i.e., the likelihood scores were all above the threshold). In particular, we can observe that the reference phones for the word /itong/ were forced-align within the span of the uttered phrase /(itoy-) itong/ without generating a score below the threshold. The second plot of Figure 10 shows the result of the word-level forced-alignment process for implementing the WDA, wherein the reference text consists of the three words /TINATAWAG/,/DIN/, and /ITONG/. Note that the second plot shows the same input speech as that of the first plot. The word boundaries shown in the second plot are system-generated and are used by the system to calculate the word durations. We may observe that the reference word /ITONG/ was forced-aligned within the span of the uttered phrase /(itoy-) itong/ which contain a miscue. Consequently, the detected duration of the word /ITONG/ became significantly higher than normal. The normal range for word duration is based on measurements made from the good pronunciations in the CFSC. The third plot in Figure 10 shows the result of word-level alignment process for an input speech corresponding to a good pronunciation. A significant difference may be observed when the detected relative word duration for the word /ITONG/ in the second plot is compared to that in the third plot. In fact, the system was able to detect that the word /ITONG/ from the input speech, shown in the second plot, has a 136% deviation from the normal. Since a 136% deviation is above the threshold set for the system, the reading miscue is therefore detected by the WDA method.

The effectiveness of the WDA method in detecting reading miscues in Filipino may further be explained by two main reasons: (1) the suitability of the WDA method to the nature of reading miscues in children's Filipino read speech; and, (2) the suitability of the WDA method to the nature of the Filipino language.

Table 2 in the previous section has listed the different types of reading miscues/ disfluencies found in children's Filipino read speech taken from the CFSC as: partial

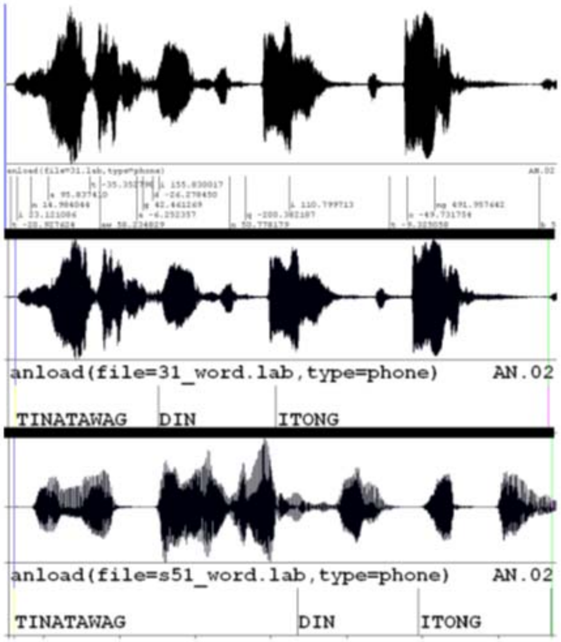


Figure 10. A specific example of a case wherein a reading miscue, undetected by the phone-level reference verification (RV) method, was detected by the word duration analysis (WDA) method.

word; hesitation pause; insertion; repetition; substitution; deletion; and, elongation. An examination of the nature of these reading miscues would show that all of them, except substitution, are time-dependent. That is, these reading miscues would affect the effective word durations as measured by the system. Moreover, the authors have observed from the test set that hesitation pauses, partial words (mostly followed by a short pause or a restart), and repetitions are the miscue types that usually cause the largest word duration deviations. Since the aforementioned miscue types constitute the majority of the miscues found in the test set, the WDA method therefore generally becomes an effective way of detecting the reading miscues.

Since Filipino is a syllable-timed language, the insertion or deletion of syllables or words, as well as pauses, definitely affects the effective word durations, as measured by the RMD system. The WDA method for the RMD may therefore be shown to be especially effective for syllable-timed languages, such as Filipino.

DESIGN OF THE AUTOMATED READING TUTOR FOR FILIPINO

The ART for Filipino presented in this paper has the following major components: (1) the RMD; (2) the oral/visual feedback and instruction set; and, (3) the graphical user interface.

The RMD for Filipino, which is the core of the ART, employs a two-level RV-plus-WDA architecture as presented in the previous section. The performance measures for the RMD are the FA rate and the MDerr rate. The best result of the offline performance evaluation tests shows that the RMD's operating point may be calibrated, such that the FA rate is approximately 3% while the MDerr rate is approximately 5%. The FA and MDerr rates that the authors obtained for the RMD for Filipino prove that it is at par with the state-of-the-art RMDs (Duchateau et al. 2006; Liu et al. 2008; Black et al. 2011) reported in the literature. For comparison, Table 5 summarizes the performance of the two-level RV-plus-WDA-based RMD presented in this paper, together with those from other systems reported in the literature.

The oral/visual feedback and instruction set allows active interaction between the machine and the child (user or learner). The ORF instruction set used in this study is a set of pre-recorded speech of a human tutor, who is a reading expert and an education sector research collaborator from the College of Education at the University of the Philippines Diliman. Any desired sentence within the instruction set can automatically be played-back by the ART system whenever there is a need to model the correct pronunciations of the words in the text passages.

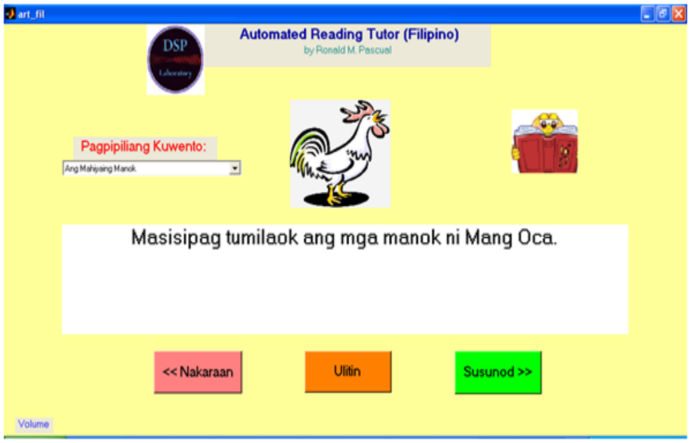
Table 5. Summary of specifications of various state-of-the-art reading miscue detectors (RMD)

State-of-the-art RMDs	False Alarm (FA) rate	Miscue Detection Error (MDerr) rate	Language
Black et al. (2011)	8.1%	11.5%	English
Liu et al. (2008)	5.82%	9.07%	Chinese Mandarin
Duchateau et al. (2006)	2.1% - 8.4%	23.1% - 16.9%	Dutch
Two-level RV-plus-WDA	3%	5%	Filipino

As briefly discussed in the previous section, nine Filipino text passages were used for both the speech database collection and the ART system development. The six short stories were provided through a non-disclosure agreement by Adarna Publishing House, a leading publisher of Filipino short stories for children. The three expository texts were adapted from various grade school textbooks used in the Philippines. Each of the text passages adapted was provided with a corresponding reading age recommendation by its publisher. Moreover, the text passages have been selected through the suggestions of research collaborators. The nine text passages all together have a total of 2,169 words (about 650 of which are unique) and 290 sentences.

The ART also provides audible comments and visual animations as positive feedback or "praise" (Mostow and Aist 1999) in response to a perceived good reading performance of the learner. According to suggestions in the literature, giving positive feedback is a powerful motivation and it demonstrates that the ART system is a perceptive and responsive audience for the learner's efforts. In the ART presented in this study, the authors employed the audible comments: "Mahusay!" (Excellent!), "Magaling!" (Good!), and "Kahanga-hanga!" (Admirable!). The positive comments are alternately played-back by the system whenever no reading miscues were detected from the input speech. Aside from audible comments, the ART also displays an animated icon whenever the aforementioned comments are being played-back.

Figure 11 shows the graphical user interface of the ART in Filipino. The graphical user interface of the ART allows the reader to select a story and navigate through the sentences within the selected story. An important design consideration for the interface is its simplicity because the intended user may be as young as a Grade 1 student (typically aged 5 to 7).



Pagpipiliang Kuwento: Ang Mahiyaing Manok

Masisipag tumilaok ang mga manok ni Mang Oca.

Figure 11. The graphical user interface of the automated reading tutor (ART) for Filipino.

AUTOMATED READING TUTOR (ART) INTERVENTION PROGRAM: A PILOT STUDY

Design of the ART Intervention Program

The ART intervention program or the pilot study, a pioneering experiment in computer-assisted ORF instruction in Filipino, is basically a reading tutorial program that makes use of the ART presented in the previous section. The within-subjects experiment involved a group of six grade-2 students from the University of the Philippines Integrated School (UPIS), and consists of two separate one-month periods. During the first period of the experiment, the experiment group depended only on regular classroom instruction for improving their reading skills. During the second period, the ART was used by the experiment group, in addition to the regular classroom instruction. The ART intervention program allowed the group to use the ART for about 45 minutes per day, three days per week. In order to evaluate the effectiveness of the ART in providing reading skill improvement to the students, three sets of oral reading fluency assessments (ORFA) were administered to the experiment group. Figure 12 graphically summarizes the pilot study experiment and its timeline. As we can verify from Figure 12, the first ORFA was given prior to the start of period 1, the second ORFA at the end of period 1, and the third ORFA at the end of period 2.

All three ORF assessments were administered by an expert, a Filipino teacher who also has a research experience in automated Filipino essay evaluation. As suggested by research collaborators, the ORFA employed both familiar and unfamiliar text passages, in order to obtain a more complete observation regarding the effects of the use of the ART in student's learning for both text structures. The ORFA also included a comprehension exam that has an objective-type and an essay-type components. In this study, the authors employed the following three commonly used ORFA measures in the literature: (1) number of words correct per minute (WCPM), (2) 16-point multi-dimensional ORF score; and, (3) reading comprehension scores. Group gain score analysis, also known as difference score analysis (Smolkowski 2013), was selected because it provides a simple and precise, yet unbiased and reliable way of interpreting the true change (Ragosa 1983). For instance, the WCPM gain scores clearly and meaningfully tell educators whether the experiment group improved, retained, or deteriorated, and by precisely how much, in their reading skills (Smolkowski 2013).

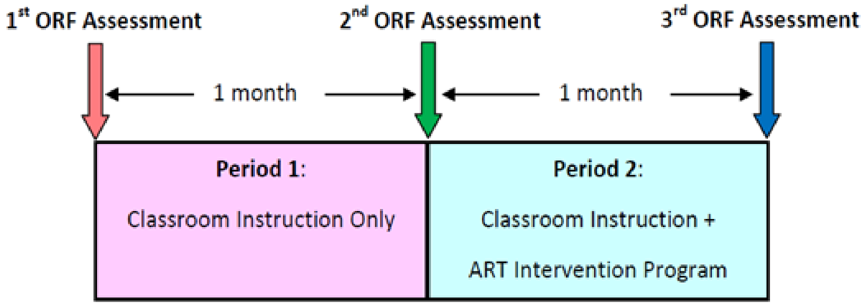


Figure 12. Pilot study experiment and timeline.

Results of the ART Intervention Program and the Oral Reading Fluency (ORF) Assessments

The primary result of the ART Intervention program presented in this study is expressed in terms of WCPM, the most widely used measure for ORFA (Rasinski 2004). Figure 13 summarizes the average or group WCPM improvements for all the students in the experiment group. An important pattern that may be noted in Figure 13 is that the group WCPM slope for the second period became abruptly higher compared to the slope for the first period. Thus, there was a significantly higher improvement in group WCPM in period 2 than that in period 1. It therefore suggests that, on the average, the reading tutor had a positive effect of accelerating the improvement of the students' ORF.

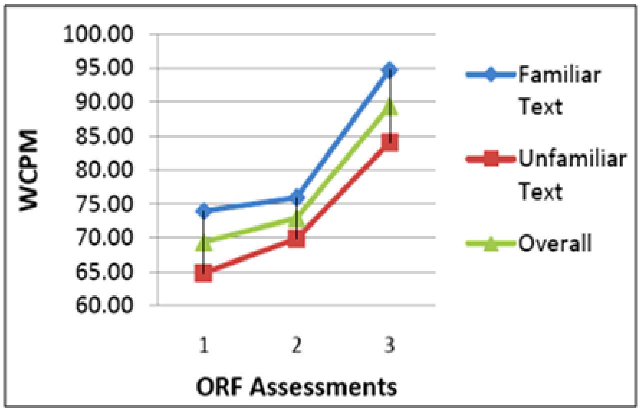


Figure 13. Group average number of words correct per minute (WCPM) for the three oral reading fluency assessments (ORFA).

In Table 6, the significantly higher overall average WCPM gain for period 2 compared to that for period 1 clearly shows that the experiment group significantly improved their ORF after using the ART. The normal growth (due to the usual classroom instruction) of the group in reading fluency for one month is shown in Table 6 to be only 3.53 words per minute. After using the ART for a month however, the improvement of the group suddenly jumped up to 16.46 words per minute, an improvement which is 12.93 words per minute higher than the normal.

To have a better idea on the magnitude of improvement in the reading fluency of the group due to the use of the ART, we may calculate the WCPM gain ratio (i.e., group gain for period 2, divided by the group gain for period 1). The overall WCPM gain ratio was calculated to be 4.66. This means that the fluency improvement rate in period 2 (i.e., when the ART was used) improved by 466% than the normal

Table 6. Words correct per minute (WCPM) group gains for the two experiment periods

	Group Gain (WCPM)		
	Familiar Text	**Unfamiliar Text**	**Familiar Text**
Period 1 (Without using the ART)	2.01	5.05	**3.53** (*SD=3.88*)
Period 2 (Using the ART)	18.75	14.17	**16.46** (*SD=8.56*)

Note. SD = Standard Deviation

learning rate. In other words, after the ART was used for a month, the experiment showed that the reading fluency of the group has been accelerated by an amount of time roughly equivalent to three and a half months.

In order to show that the larger WCPM group gain in period 2 is indeed attributable to the use of the ART, the authors calculated the correlations between score gains in period 1, the score gains in period 2, and the score gains in the whole experiment period. The correlation of the score gains in period 2 and the score gains in the whole experimental period (i.e., periods 1 and 2 combined) shows a strong and significant relationship (i.e., with a coefficient of 91.17% at $p < 0.00005$) between the score gains in period 2 and the whole two-month experimental period. Therefore, the observed overall improvement of the students in their reading fluency can indeed be attributed to the use of the ART.

To further illustrate that it is less likely that the reading fluency improvement of the students in the experiment group is due to their normal learning rate trends or to random chance, the authors also referred to the ORF norms used for English (Hasbrouck 2006). In doing this, the authors emphasize that their purpose is not to directly compare the WCPM levels observed from their experiment to the normal WCPM levels in English. The normal WCPM levels for English are expected to be different from normal WCPM levels in Filipino due to the differences in features and orthography between the two languages. Thus, the authors only referred to the ORF norms in English for the purpose of generally comparing their experimentally observed reading fluency improvement trend to the expected normal reading fluency trend in English. The aforementioned WCPM trend comparison for Grade 2 level is given in Figure 14. A simple graphical analysis on the WCPM trends shown

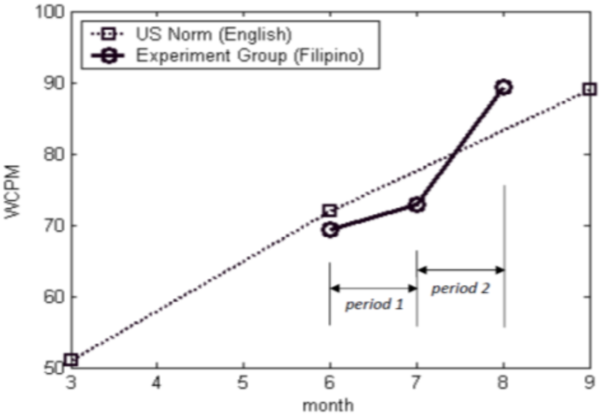


Figure 14. Comparison of group average number of words correct per minute (WCPM) trends between US norm (English) and experiment observations (Filipino).

in Figure 14 would suggest that the reading fluency improvement observed from the experiment group during period 2 is high relative to English ORF norm, and significantly higher than the reading fluency improvement observed during period 1.

Moreover, the expected English ORF improvement throughout the entire school year is fairly linear. By contrast, the reading fluency improvement trend observed from the experiment group for the entire two-month experiment period highly deviated from the linear trend. Thus, simple trend analysis clearly shows that the ORF improvement of the students in the experiment group during period 2 is unusually high, and this may be attributed to the treatment made during the period (i.e., the use of the ART).The second set of ORFA results is based on the measure known as the ORF-16 score obtained through the use of a 16-point multi-dimensional ORF rubric proposed by Rasinski (2004). The four dimensions considered in the ORF-16 rubric are: (1) expression and volume; (2) phrasing; (3) smoothness; and, (4) pace.

The trend in the ORF-16 group scores shown in Figure 15 also seem to agree with that of the WCPM group scores presented earlier in this section. We may observe from Figure 15 that there was a sudden increase in the reading fluency of the students in the experiment group during period 2, the period when the ART was used by the group. Thus, in a similar way that was shown earlier in this section, it follows that the sudden improvement in the student's reading fluency may be attributed to the use of the ART. The overall ORF-16 group gain ratio, which is calculated to be 6.0, means that the observed reading fluency improvement of the students in the experiment group during the time that they were using the ART is about six times better compared to the time that they were not using the ART.

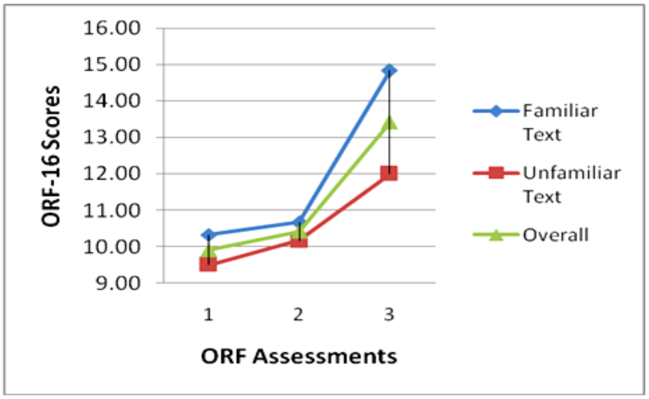


Figure 15. Sixteen-point multi-dimensional oral reading fluency (ORF-16) group scores for the three oral reading fluency assessments (ORFAs).

In order to see how the reading fluency development has affected the comprehension of the students in the experiment group, the third set of ORFA results was based on the comprehension exam scores. The plots in Figure 16 show that, on the average, the student's comprehension also abruptly improved during period 2, the time when they were using the ART. The overall comprehension exam score gain ratio, computed to be 4.43, indicates that the students have improved in their comprehension by more than four times after using the ART.

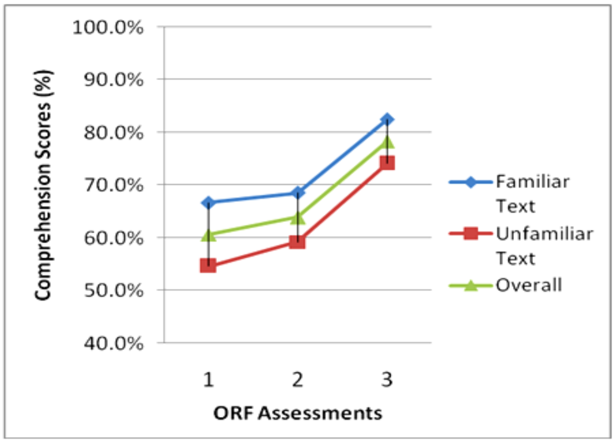


Figure 16. Comprehension exam group scores for the three oral reading fluency assessments (ORFAs).

CONCLUSION AND FUTURE DIRECTIONS

In this paper, the authors presented a two-level RMD for the design of an ART for Filipino that uses phone-level RV and WDA methods. The results of offline tests showed that the RMD's performance (i.e., false alarm rate ≈ 3% and misdetection rate ≈ 5%) is at par with those from state-of-the-art RMDs (Duchateau et al. 2006; Liu et al. 2008; Black et al. 2011) reported in the literature. The advantages of the RV-plus-WDA RMD are design simplicity (i.e., it did not require building a complex language model or using an adaptation technique) and low training cost (i.e., it only required 1.5 hours of training data).

The authors of this study have discussed the design of the ART prototype that integrates the two-level RV-plus-WDA RMD, the user interface, and the feedback and instruction sets. The authors suggest the following design considerations: (1) user interface simplicity on account of very young users; (2) minimal interventions to avoid children's frustration; and, (3) the use of positive feedback or "praise" that most children seem to appreciate. Moreover, it is suggested that all model pronunciations in the instruction set should contain the "correct" or acceptable prosodic features because it has been observed that students have the tendency to adopt these features.

The authors have presented in this paper an experimental procedure for evaluating the effectiveness of an ART for Filipino that involves an ART intervention program and a set of ORFA for a small group of students. The results of the ART Intervention experiment clearly showed the ART's effectiveness in improving the students' ORF in terms of WCPM, ORF-16 scores, and comprehension scores. Specifically, the results of the ORFAs showed that, after using the ART, the students, on the average, have improved in their WCPM by 4.66 times compared to the period when they were not using the ART. Correlation and trend analysis undoubtedly implies that the improvement of the students in their reading fluency was indeed attributable to the use of the ART. Similarly, after using the ART, the students have improved in their ORF-16 scores by 6.0 times compared to the period when they were not using the system. The results of experiment have also shown that, after using the ART, the students, on the average, were 4.4 times better in reading comprehension than when they were not using the ART.

With all the positive results that the authors obtained from the study, the ART in Filipino seems to be a promising and important Filipino speech technology to further develop and implement for the primary education system in the Philippines. Future directions for this study include the development of an automated oral reading assessment system for children's Filipino speech, adaptation of the RMD system for nonnative speakers of Filipino (or those who speak Filipino as their second or third language), and adaptation of the system design methods for other Philippine languages and other related applications.

ACKNOWLEDGMENTS

The authors would like acknowledge the UP Digital Signal Processing Laboratory, UP Integrated School, UP College of Education, UP Department of Linguistics, and Adarna House for lending their support to the study. This research was funded by the CHEDSEGS grant from the Commission on Higher Education of the Philippines, in collaboration with the Office of the Vice-Chancellor for Research and Development of University of the Philippines Diliman.

REFERENCES

Batliner A, Blomberg M, D'Arcy S, Elenius D, Giuliani D, Gerosa M, Hacker C, Russell M, Steidl S, Wong M. 2005. The PF_STAR children's speech corpus. In: Proceedings of INTERSPEECH; Lisbon, Portugal. p. 2761-2764.

Black M, Tepperman J, Narayanan S. 2011. Automatic prediction of children's reading ability for high-level literacy assessment. IEEE Trans. on Audio, Speech and Language Processing. 19(4):1015-1028.

Bolaños D, Ward W, Cole R. 2009. A reference verification framework and its application to a children's speech reading tracker. In: Proceedings of 2nd Workshop on Child, Computer and Interaction; NY, USA: ACM. p. 22.

Cleuren L, Duchateau J, Ghesquiere P, Van hamme H. 2008. Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In: Proceedings of 6th International Conference on Language Resources and Evaluation; Morocco.

Duchateau J, Wigham M, Demuynck K, Van hamme H. 2006. A flexible recogniser architecture in a reading tutor for children. In: Proceedings of ITRW on Speech Recognition and Intrinsic Variation; Toulouse, France.

Duchateau J, Kong Y, Cleuren L, Latacz L, Roelens J, Samir A, Demuynck K, Ghesquiere P, Verhelst W, Van hamme H. 2009. Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Communication. 51(10):985-994.

Duong M, Mostow J, Sitaram S. 2011. Two methods for assessing oral reading prosody. ACM Trans. on Speech and Language Processing. 7(11):14.

Gao J, Li A, Xiong Z. 2012. Mandarin multimedia child speech corpus: Cass_Child. In: Proceedings of 2012 International Conference on Speech Database and Assessments (Oriental COCOSDA); IEEE Xplore.

Gerosa M, Giuliani D, Narayanan S, Potamianos A. 2009. A review of ASR technologies for children's speech. In: Proceedings of 2nd Workshop on Child, Computer and Interaction; Cambridge, MA, USA: ACM. p. 7.

Guevara RC, Garcia I, Santos T, Nolasco R. 2010. A computational approach to Filipino speech rhythm. In: Proceedings of 1st Philippine Conference-Workshop on Mother Tongue-Based Multilingual Education; Cagayan de Oro City, Philippines.

Hagen A, Pellom B, Cole R. 2003. Children's speech recognition with application to interactive books and tutors. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding; St. Thomas, Virgin Islands: IEEE Xplore. p. 186-191.

Hasbrouck J. 2006. Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher. 59(7):636-644.

Kazemzadeh A, You H, Iseli M, Jones B, Cui X, Heritage M, Price P, Anderson E, Narayanan S, Alwan A. 2005. TBALL data collection: The making of a young children's speech corpus. In: Proceedings of Interspeech; Lisbon, Portugal. p. 1581-1584.

Liu C, Pan F, Ge F, Dong B, Zhao Q, Yan Y. 2008. Application of LVCSR to the detection of Chinese Mandarin reading miscues. In: Proceedings of 4th International Conference on Natural Computation; Jinan, China: IEEE Xplore. p. 447-451.

Mills-Tettey G, Mostow J, Dias MB, Sweet T, Belousov S, Dias MF. 2009. Improving child literacy in Africa: Experiments with an Automated Reading Tutor. In: Proceedings of 3rd International Conference on Information and Communication Technologies and Development; Doha, Qatar: IEEE Xplore. p. 129-138.

Mostow J, Roth S, Hauptmann A, Kane M. 1994. A prototype reading coach that listens. In: Proceedings of 12th National Conference on Artificial Intelligence; Seattle, WA: ACM. p. 785-792.

Mostow J, Aist G. 1999. Giving help and praise in a reading tutor with imperfect listening - Because automated speech recognition means never being able to say you're certain. The Computer Assisted Language Instruction Consortium (CALICO) Journal. 16(3):407-424.

Mostow J, Nelson-Taylor J, Beck J. 2013. Computer guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens. Journal of Educational Computing Research. 49(2):249-276.

[NRP] National Reading Panel (US). 2000. Teaching children to read. Panel Report issued for the National Institute of Child Health and Human Development, U.S. Department of Health and Human Services. NIH Pub. No. 00-4769.

Pascual R, Guevara RC. 2012a. Developing a children's Filipino speech corpus for application in automatic detection of reading miscues and disfluencies. In: Proceedings of IEEE TENCON 2012: IEEE Asia Pacific Region International Conference; 2012; Cebu City, Philippines: IEEE Xplore.

Pascual R, Guevara RC. 2012b. Developing an automated reading tutor in Filipino for primary students. In: Proceedings of 2nd Philippine Conference-Workshop on Mother Tongue-Based Multilingual Education; Iloilo City, Philippines.

Ragosa D. 1983. Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement. 20(4):335-343.

Rahman F, Mohamed N, Mustafa M, Salim S. 2014. Automatic speech recognition system for Malay speaking children. In: Proceedings of the 2014 Third ICT-ISPC; Nakhon Pathom, Thailand: IEEE Xplore. p. 79-82.

Rasinski T. 2004. Assessing reading fluency. Hawaii: Pacific Resources for Education and Learning. p. 1-25.

Rayner M, Tsourakis N, Baur C, Bouillon P, Gerlach J. 2014. CALL-SLT: A spoken CALL system based on grammar and speech recognition. Linguistic Issues in Language Technology. 10(2):1-23.

Reeder K, Shapiro J, Wakefield J, D'Silva R. 2015. Speech recognition software contributes to reading development for young learners of English. International Journal of Computer-Assisted Language Learning and Teaching. 5(3):60-74.

Russell M. 2010. Speech technologies for children. New Orleans: IEEE Signal Processing Society - STLC Newsletter.

Smolkowski K. [Internet]. 2013. Gain Score Analysis. Oregon: Oregon Research Institute; [cited 2016 Dec]. Available from http://homes.ori.org/keiths/Tips/Stats_GainScores.html.

Tsau S. 2012. The effects of an automatic speech analysis system on enhancing EFL learners' oral reading fluency. Procedia-Social and Behavioral Sciences. 64(2012):141-150.

Wasik B, Slavin R. 1993. Preventing early reading failure with one-to-one tutoring: A review of five programs. Reading Research Quarterly. 28(2):178-200.

Young S, Evermann G, Gales M, Woodland P. 2006. The HTK Book [Internet]. UK: Cambridge University Engineering Department; [cited 2010 Nov 19]. Available from http://htk.eng.cam.ac.uk.

____

Dr. Ronald M. Pascual <ronaldmpascual@gmail.com> is an Associate Professor and Assistant Director of the Electronics and Electrical Engineering Department of FEU Institute of Technology, Manila. He received his Ph.D. in Electrical and Electronics Engineering from University of the Philippines Diliman as a CHED scholar, his M.S. in Electronics and Communications Engineering from De La Salle University, Manila as a DOST scholar, and his B.S. in Electronics and Communications Engineering from Pamantasan ng Lungsod ng Maynila. His research interests include speech signal processing, and speech technology development.

Dr. Rowena Cristina L. Guevara is the Undersecretary for Research and Development of the Department of Science and Technology (DOST) and a Professor of the Digital Signal Processing Laboratory of the University of the Philippines Diliman. She was a former Executive Director of DOST-Philippine Council for Industry, Energy, and Emerging Technology Research and Development, and was a former Dean of the College of Engineering of the University of the Philippines Diliman. She received her Ph.D. in Electrical Engineering from University of Michigan, Ann Arbor as a DOST scholar, and her M.S. and B.S. in Electrical Engineering from the University of the Philippines Diliman. Her research interests include speech signal processing, and audio and communications signal processing.