A Supervised Phrase Selection Strategy for Phonetically Balanced Standard Yorùbá Corpus

Adeyanju Sosimi[1,*], Tunde Adegbola[2], and Omotayo Fakinlede[1]

[1]University of Lagos, Akoka-Lagos, Nigeria
azeezadeyanju@gmail.com, oafak@unilag.edu.ng
[2]Africa Language Technology Initiative, Bodija, Ibadan, Nigeria
taintransit@hotmail.com

Abstract. This paper presents a scheme for the development of speech corpus for Standard Yorùbá (SY). The problem herein is the non-availability of phonetically balanced corpus in most resource-scarce languages such as SY. The proposed solution herein is hinged on the development and implementation of a supervised phrase selection using Rule-Based Corpus Optimization Model (RBCOM) to obtain phonetically balanced SY corpus. This was in turn compared with the random phrase selection procedure. The concept of Exploitative Data Analysis (EDA), which is premised on frequency distribution models, was further deployed to evaluate the distribution of allophones of selected phrases. The goodness of fit of the frequency distributions was studied using: Kolmogorov Smirnov, Andersen Darling and Chi-Squared tests while comparative studies were respectively carried out among other techniques. The sample skewness result was used to establish the normality behavior of the data. The results obtained confirmed the efficacy of the supervised phrase selection against the random phrase selection.

Keywords: Standard Yorùbá, Corpus, Rule-Based Corpus Optimization Model, Phrase selection, Automatic Speech Recognition.

1 Introduction

Human Language Technology (HLT) applications currently exist for a vast majority of languages of the industrialized nations but this is not the case with most African languages such as the Standard Yoruba (SY) language. One of the impediments to such development for these Africa languages is the relatively lack of language corpus. Unlike the case of most African languages, language corpus have over the years been fully developed other continents as seen in the Thai language, Chinese and French languages amongst others which have different corpus from the English language. HLT development is premised on availability of phonetically rich and balanced digital speech corpus of the target language [1]. However, only few African languages have speech corpus required for HLT development, which in recent times have been given a higher priority as a result of global technological influence. SY is one of the few indigenous Nigerian languages that have benefitted from the platform of HLT.

A. Gelbukh (Ed.): CICLing 2015, Part II, LNCS 9042, pp. 565–582, 2015.
DOI: 10.1007/978-3-319-18117-2_42

[2] reported that SY is the native language of more than 30 million people within and outside Nigeria. With respect to speech recognition, the Yorùbá language bears a challenging characteristic in the usage of tones to discriminate meaning. Yorùbá speech is constructed by appropriate combination of elements from a phonological alphabet of three lexically contractive vocal gestures, namely consonants, vowels, and tones. According to [3], the three distinct tones, therefore, widen the scope of an x- syllable Yorùbá word to 3^x possible permutations while [4] considered two systems for SY ASR: oral vowels using fuzzy logic (FL) and artificial neural network (ANN) based models. [5] considered additional feature extraction methods to evaluate the effect of voice activity detection in an isolated Yorùbá word recognition system. However, in the case of continuous speech, the signal is affected by many factors such as sentence prosody, co-articulation, speaker's emotion, gesticulation, etc. [6]

To accomplish the target for the SY language, there is a need for efficient and effective continuous speech corpus development as presented by [7], where it was stated that the construction of phonetically rich and balanced speech corpus is based on the selection of a set of phrases. In the literature, various techniques have been proposed for such selection, and a major constraint with this is the cost development. An approach for such selection is to ensure that allophone interaction and distribution of phrases have equal parity without loss of information and also, not undermining language syntactic rules. [8] reported that uniform distribution and frequency of occurrence of phones appears to be the dominant paradigm in assessing allophones optimality. The adopted strategy for new language resource development is dependent on the task; the first scheme is particularly used for training a Text-to-Speech (TTS) system while the second type is better adapted for the development of ASR systems [8]. Generally, optimal distribution of allophones is of significance when developing corpora for resource-scarce language [1].

[9] reported that the use of prompts produces quality corpus for training the ASR system. [10] discussed the design of three large databases to cope with the challenges of the ASR system. [8], [11], [12], [13], and [14] considered the design and development of phonetically balanced corpus for an under-resourced language. These authors premised their work on the understanding that corpus are word- and sentence-based. However, corpus development using the sentence or phrase selection technique presents some challenges, which include how to harvest the selected phrase in the target language, how to preserve context integrity and how to classify a phonetically rich and balanced phrase [15].

This research seeks to address the challenges facing the SY corpus development. The focus is to develop and implement a Supervised Prompt Selection Strategy for the generation of a phonetically balanced SY speech corpus. Furthermore is the integration of a tone explicit model for addressing the tone characteristics for SY and test a speech recognizer using the SY corpus.

The following sections of this paper includes: Section 2, which presents the proposed methodology and implementation procedures while section 3 presents the evaluation SY speaker independent ASR based on the RBCOM corpus. Finally, Section 4 gives a brief conclusion and likely extensions required to enhance the performance of the Rule-Based Corpus Optimization Model (RBCOM).

2 Methodology

2.1 SY Corpus Development

The methodology of the SY corpus development was carried out based on the overall system design as illustrated in Figure 1. Once the text data is captured, a formatting process is carried out after which an optional set of text are generated followed by selection and evaluation of n-gram after which a prompt selection process is effected.

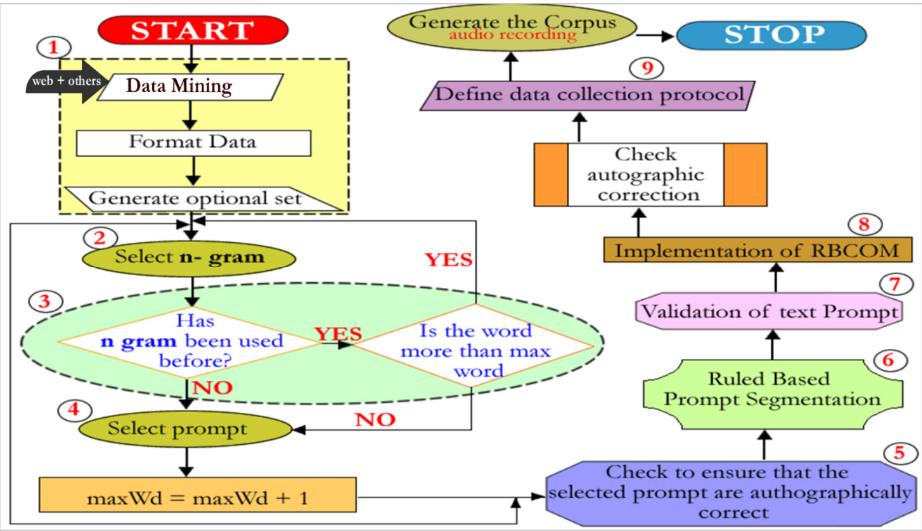

Fig. 1. Corpus Development Approach

2.2 Rule-Based Prompt Segmentation

The rules utilized herein, are adapted from the existing SY literature. Rule-based prompt segmentation was integrated into the general approach for pragmatic evaluation of phone interaction with a view to having optimal coverage. The RBCOM is based on the three possible syllable structures of SY; these include consonant-vowel CV, vowel V and syllabic nasal N. The rules for Yorùbá Phoneme Interaction are as follows:

Phonemes (Consonants)

1. The following phonemes co-occur with any other phonemes at the word initial position, middle, or final word position to have a CV syllable b, t, d, j, k, g, p, gb, f, s, ṣ and m.
2. 'h', 'r', 'l', 'y' and 'w' co-occur with oral and nasalized vowels (phonemes). Any of the consonants come before any of the vowels.
3. The consonant 'n' co-occurs with oral vowels (phonemes). It comes before any of the oral vowels in order to form a CV or CVV syllable structure.

Phonemes (Oral Vowels)

1. All the oral vowels and three nasalized vowels '-in-', '-ẹn-' and '-un-' co-occur with any other phonemes at the word initial, middle and final positions to form VV, CV or VC syllable structure.
2. The nasalized vowels '-an-' and '-ẹn-' co-occur with some consonants (phonemes) to form a CV syllabic structure. The vowel follows the consonants.

These rule sets were used for the segmentation of SY corpus into basic syllabic units. In this work, each valid syllable constitutes a class, and members of each class were considered as objects. Since a specific set of prompts were given to the respondent, this work viewed those set of prompts as documents for the purpose of clarity and understanding. The number of documents depends on the anticipated number of respondents. Hence, for each respondent, there exists a document containing finite lines of sentences for which corresponding audio files will be created. To have an optimal coverage of phones, RBCOM will be implemented.

2.3 Implementation of RBCOM

The implementation of the Rule Based Corpus Optimization Model is presented in this sub-section. This has its link from item eight as shown in figure 1. In summary, the algorithms as presented herein are focused at developing a phonetically balanced corpus for a resource scarce language like SY.

2.3.1 Algorithm 1: Random Selection of Speech Document in a Pool for Analysis

```
Algorithm 1
For  k = k to K;  k = 1,...,K          /*            k = document
For  i = 1 to I;  i = 1,...,N_T        /*            N_T = Total No of sentence in R
                                       /* R = Respository
Return  Rnd              /*return a random number between 0 and 1
f_r = [Rnd * i]
Select  S_ik = ⌈f_r⌉
Let sentence r be sentence  i in document k
where [r ∈ f_r]
Next i
Next k
```

Section 2.3.1 presents the psuedocode of Algorithm 1 whose task is to select sentences from the repository pool for document composition. To determine sentence $\boldsymbol{i}$ in a document $\boldsymbol{k}$, a random number $\boldsymbol{Rnd}$ is generated $\ni \mathbf{0} \leq \boldsymbol{Rnd} \leq \mathbf{1}$ and multiplied by $\boldsymbol{N_T}$. Here, $\boldsymbol{S_{ik}} = \lceil \boldsymbol{f_r} \rceil$ i.e. sentence $\boldsymbol{i}$ in a document $\boldsymbol{k}$. This selection process continues until the finite number of lines of sentences are achieved for the respondents $\ni \mathbf{1} \leq \boldsymbol{k} \leq \boldsymbol{K}$. Sentence*s* $\boldsymbol{i}$ as contained in $\boldsymbol{k}$ are then analyzed based on the algorithmic design described in Algorithm 2.

2.3.2 Algorithm 2: Population of Investigatory Array

```
Algorithm 2
Read {D_k}      k = 1, ..., K              /* create an array of for documents
 Read {S_nk}     n = 1, ..., N              /* S_nk = sentences n in documents k
ead {L_j}  j = 1,2, ... ..., J              /* syllables

 Do while k ≤ K                             /* read each document
   k = k + 1
   For  n = 1 to N                          /* read each sentence in a document
   n = n + 1
  Read S_nk                                 /* read sentence n in document k
   Flag = false
   Do while Flag = false for j = 1 to J     /* for syllable types
     For each L ϵ S_nk                      /* Read syllable in sentence
     if L ϵ C_j                             /* is syllable in class j of syllables?
     C_jk = C_jk + 1                        /* count j in S_nk  and store
  else
       C_jk = C_jk                          /* syllable occurrence frequency
      Next j
        Test_ End_ of Sentence S_nk
     If End_of Sentence S_nk
       Flag = True
  Loop
    Next n
      Loop
```

In Algorithm 2, the investigatory array is populated based on SY syllable space as obtained in **Section 2.2**. For each document $\boldsymbol{k}$, select and read sentence $\boldsymbol{n}$ where $\boldsymbol{n} = 1,2, \dots \dots \boldsymbol{N}$. Furthermore, within each sentence $\boldsymbol{n}$, read syllable, classify and count number of syllable classes $\boldsymbol{j}$ where $\boldsymbol{j} = \mathbf{1}, \mathbf{2}, \dots \dots \boldsymbol{J}$ as $\boldsymbol{C_{jk}}$ until the search gets to the end of the sentence. Thus, $\{\boldsymbol{C_{jk}}\}$ for syllable class $\boldsymbol{j} = \mathbf{1}, \mathbf{2}, \dots \dots \boldsymbol{J}$ and document $\boldsymbol{k} = \mathbf{1}, \mathbf{2}, \dots \dots \ \boldsymbol{K}$ constitute an array $\boldsymbol{J} \times \boldsymbol{K}$, with non-negative entries (i.e. including zero entries where a syllable class is not found within a document) $\boldsymbol{C_{jk}}$ representing the number of times each syllable $\boldsymbol{j}$ is found in document $\boldsymbol{k}$. Syllable $\boldsymbol{j}$ with $\boldsymbol{C_{jk}} = \mathbf{0}$ are identified and strategies for the replacement of $\boldsymbol{S_{ik}}$ are described in Algorithm 3.

2.3.3 Algorithm 3: Identification of Array Entries and Replacement Strategies

Algorithm 3

Step I: Identification of array entries with zero and replacement

```
Read {C_jk}                         /* Read investigation array
Do while zero_content> 0
For j = 1 to J
For k = 1 to K
if C_jk > 0 ;
goto TRAY
elseif C_jk = 0                     /* if array entry is zero
```

Step II: Search the repository for viable replacement

```
Read Repository R
Do while m ≤ N_T            /* N_T is the total number of sentences in repository
    m = m + 1
RFlag = false
Do while RFlag = False
Read r_m ∈ R                /* read sentence m in R
    For j = 1 to J
if L_y = L_j                /* Syllable y is same as syllable in class j
        P1 = P1 + 1         /* No of syllable of type j seen in a sentence m ∈ M
else
Next j
End of Sentence = True
RFlag = True                Rflag is Repository flag
Loop
if P1 > P2
C_S = Label of Candidate_ Sentence = m         /* Replacement candidate
P2 = P1
P1 = 0
Select fc_s                 /* Sentence for replacement
Loop
Read {D_jk} ≡ {C_jk}
```

Step III: Ranking Syllables in Investigation Array

$Q = 0$ /* copy array $\{C_{ik}\}$

$for\ \ v = 1, \dots \dots, V$

$w = 1, \dots \dots, W$

$if\ d_{wk} > \ d_{vk}$

$d_{vk} = Q \quad d_{vk} = \ d_{wk}\ ;\ d_{wk} = Q$

$h_v = w$ /* keep track of original label of syllable now ranked

$g_w = v$

$Next\ w$

$Next\ v$

Step IV: Determination of sentence to replace in Investigation pool

$For\ i = 1\ to\ 20$

/* top twenty syllable will be considered as candidate for removal

$q = q + 1$

$\boldsymbol{Do\ while}\ n\ \leq N$

$n = n + 1$

$For\ z = z\ to\ Z$

$g = h_z$ /* identify the original label

$Flag = false$

Do while $Flag = False$

$\boldsymbol{Read}\ S_{nk}$ /* read sentence

$\boldsymbol{if}\ L_q \in\ S_{nk}$ /* if syllable q is in sentence n

$X_{qn} =\ X_{qn} + 1$ /* increment number of syllable q in sentence

$Next\ i$

$\boldsymbol{if}\ End\ of\ statement\ n\ then$

$Flag = True$

$\boldsymbol{Loop}$

Algorithm 3 was executed in four steps. The first step is to search each document $\boldsymbol{k}$ for syllable $\boldsymbol{j}$ and the number of occurrence of syllable $\boldsymbol{j}$ is counted. Where zero entry is found in array $\{\boldsymbol{C_{jk}}\}$, i.e. $\boldsymbol{C_{jk}} = 0$, for each sentence $\boldsymbol{m}$, where $\boldsymbol{m} = \mathbf{1}$, $\mathbf{2}$,….., $\boldsymbol{N_T}$ in the repository is read as described in step two. In step three, the occurrence of syllable $\boldsymbol{j}$ in the repository is also counted and the sentence index is noted. The sentence with highest occurrence of the syllable $\boldsymbol{j}$ with $\boldsymbol{C_{jk}} = \mathbf{0}$ and syllable $\boldsymbol{i}$ with least occurrence in document $\boldsymbol{k}$ is selected. If in a document the array content $\boldsymbol{C_{jk}}$ is zero, the repository is searched for viable replacement. The highest and least occurrence of syllable $\boldsymbol{j}$ are noted. Finally, step four is the determination sentence $\boldsymbol{S_{nk}}$ to replace. All $\boldsymbol{S_{nk}}$ were read and the count of highest occurrence of syllable $\boldsymbol{j}$ are noted. The sentence $\boldsymbol{i}$ with $\boldsymbol{j}$ space syllables having the highest number of counts is selected for replacement and the array ratings obtained from Algorithm 3. This was in-turn evaluated using Algorithm 4 as presented in Section 2.3.4.

2.3.4 Algorithm 4: Evaluation Model of Array Ratings

The evaluation procedure of both the random and RBCOM generated corpus is presented in Section 2.3.4.

Algorithm 4

```
Read {C_jk}
for j = 1,2, ……..J                /*Column
for k = 1,2, ……..K                /* Row
CTs = CTs + C_jk                  /*Aggregate of k over all j
CTj = CTs/K                       /* Average count of syllables for each document k
Next k
      CTs = 0
Next j
for k = 1,2, ……. K
for j = 1,2, ……..J
RT_k = RTs + C_jk                 /*Aggregate of j over all k
RTs = RTs/J                       /* Average count of syllables for each class j
Next j
    RTs = 0
Next k
for k = 1, ……. K
for j = 1, ……..J
Obj_Sum = Obj_Sum + [|RT_j − C_jk| + |CT_k − C_jk|]
/* Model objective function
   Next j
Next k
             P4 = M               /* M is a large number
for j = 1 to W
if
      P3 < P4
  P4 = P3
Replace array {C_jk}
   else
Next w
```

Algorithm 4 depicts the basic steps involved in the Random and RBCOM corpus evaluation. To ensure optimal syllable distribution within and across all documents for **k = 1, 2, … K**, the objective of the evaluation model as encapsulated in algorithm 4 is to minimize the difference in syllable counts within and across documents without interfering with the language syntactic rules. In this algorithm the function $\boldsymbol{f(j,k)}$ is expressed as

$$f(j,k) = \sum_{j=1}^{J} \sum_{k=1}^{K} [|RT_j - C_{jk}| + |CT_k - C_{jk}|]$$

$\boldsymbol{RT_j}$ = Aggregate of j over all document $\boldsymbol{k}$
$\boldsymbol{C_{jk}}$ = Number of counts of particular syllable j in document $\boldsymbol{k}$
$\boldsymbol{CT_k}$ = Average count of syllable for each document $\boldsymbol{k}$

The processes from Algorithms 1 to 4 are iteratively repeated and values of $\boldsymbol{f(j,k)}$ compared until convergence got to the minimum. The RBCOM text prompt obtained on the implementation of Algorithms 1 to 4 were validated, based on schemes described in Section 2.4.

2.4 Validation of Text Prompt

The proposed text prompt validation model from the SY corpus development is as presented in this sub-section: Firstly, the modeling procedure for the Explorative Data Analysis (EDA) is premised on frequency distribution models, followed by the determination of the goodness of fit of the frequency distribution anchored on: (1) Kolmogorov Smirnov, (2) Andersen Darlin and (3) Chi-Squared test criteria. The evaluation procedure for the skewness of samples is based on the moment coefficient of skewness as shown in (1), (2), (3) and (4) below to evaluate m_3 and m_2 respectively.

$$\text{skewness: } g_1 = \frac{m_3}{m_2^{3/2}} \tag{1.0}$$

$$m_3 = \sum (x - \bar{x})^3 f/n \tag{2.0}$$

$$m_2 = \sum (x - \bar{x})^2 f/n \tag{3.0}$$

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \tag{4.0}$$

Where,
$\bar{x}$ = mean,
f = frequency
n = sample size
m_3= third moment of the data set
m_2= variance
g_1= skewness
G_1 = sample skewness

3 Results and Discussion

The summary of SY corpus development is presented in this section. It covers the following: SY data and experiment, evaluation of random and RBCOM generated corpus.

3.1 Data and Experiment

The texts of this corpus were selected from various data sources which includes: newspapers, magazines, journals, books, letters, handwritten texts, movie scripts and extracts from the television. This corpus is a complete set of SY contemporary texts. The texts are about different subjects, including politics, arts, culture, economics, sports, stories, etc. The SY harvested data contains a total of 206,444 words with 5,689 distinct words.

In order to achieve maximal lexeme diversity, an n-gram prompt selection was used to generate the prompt list. At the end of evaluation stage, a selection of sentences was done based on the established protocol. The process ensures that each respondent has at least 10-15 similar prompts and 185-190 randomly selected prompts. Sentences and word were verified based on the context of language and grammar. The quality of corpora greatly affects the performance of ASR system; therefore, before recording, a syllable segment of words in prompt list was analyzed. The syllable segmentation was carried out based on the rules defined in Section 2.2.

Graphs I-V as contained in Figure 2 below represents an illustration of EDA plot for syllables in a randomly selected text prompt. In testing for normality, the text prompt was seen to be negatively skewed, with skewness value less than -1 signifying a highly skewed syllable distribution.

Furthermore, a general trend of syllable occurrence and variants of distribution based on goodness of fit is presented in Table 1**.** The results of the expert model mining of allophone occurrence indicate the Generalized Pareto (GP), Dagnum, Johnson SB and Error are the best performing models for Kolmogorov Smirnov test. For Andersen Darling test, the best performing models are Dagnum, Gen. Pareto, Wakeby and Gumbel-Max. For, Chi-Squared test, Wakeby and Dagnum appear to be the best representation. This in addition, further reinforce the skewness of the allophones.

Having established the limitations of the randomly selected prompt, the initial results for the implementation of the RBCOM are as shown in Figures 3 and 4. A state of stability was attained for the objective function from the 50th iteration upwards. The new set of data generated from the point of stability was analyzed and validated for optimality as presented in Figure 5.

Graphs VI-IX as contained in Figure 5 represents an illustration of EDA plot for syllables obtained from the point of stability of the RBCOM generated prompt. The output as seen in the Figure 5 shows that the data normality and skewness is between -1 and +1 which represents an improvement in the distribution of allophones.

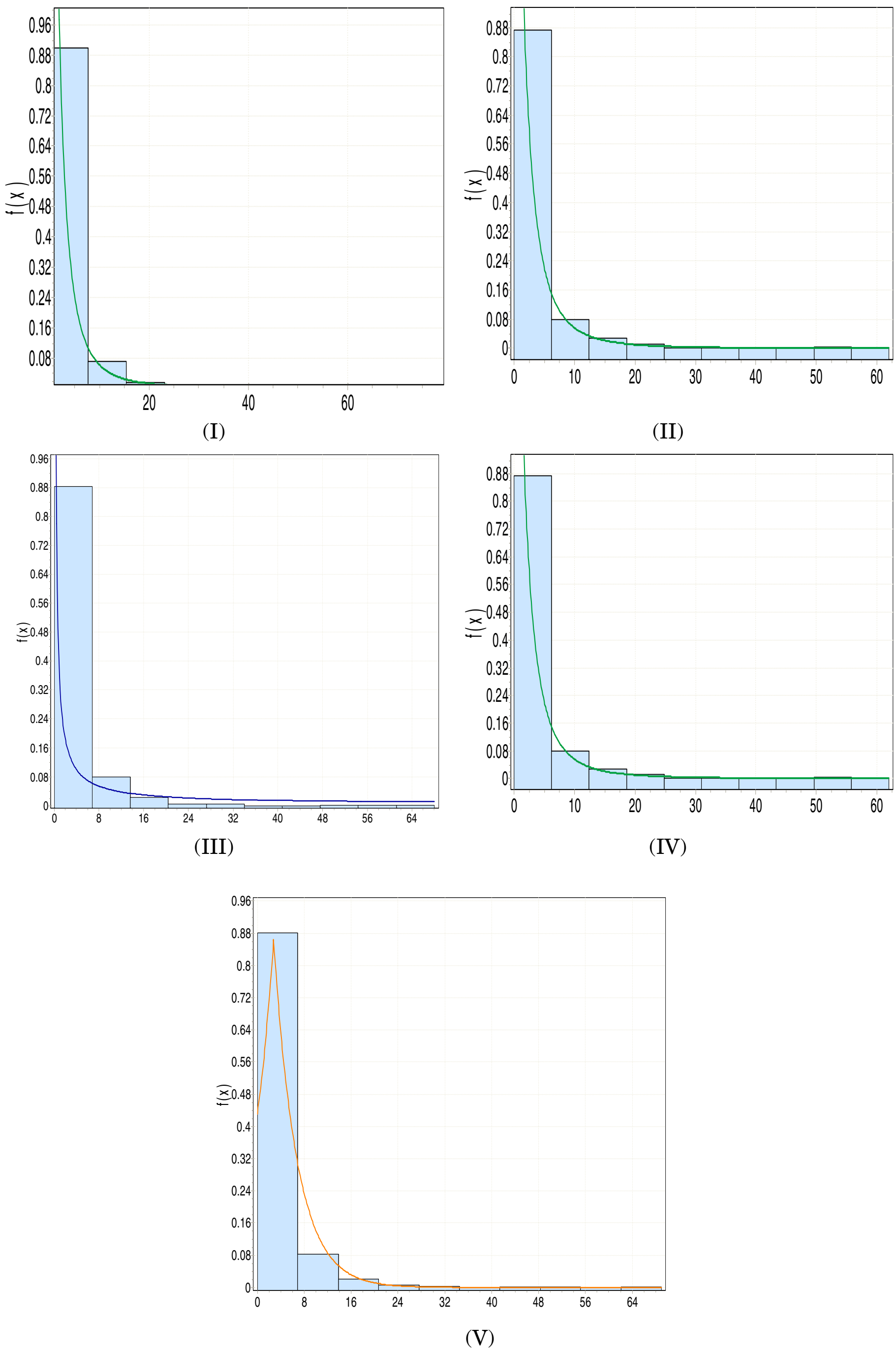

Fig. 2. SY syllable frequency trend for randomly selected prompts

Table 1. Expect Model Mining of Allophones for Random corpus

S/No	Distribution	Kolmogorov Smirnov	Andersen Darling	Chi-Squared
		Rank	Rank	Rank
TYPE I	Gen. Pareto	1	2	2
	Dagnum	28	1	33
	Wakeby	2	3	1
TYPE II	Johnson SB	1	48	NA
	Gen. Pareto	3	1	36
	Dagnum	2	8	1
TYPE III	Gen. Pareto	1	10	3
	Kumaraswamy	27	1	35
	Dagnum	13	22	1
TYPE IV	Johnson SB	1	49	NA
	Wakeby	2	1	1
TYPE V	Error	1	19	11
	Gumbel. Max	34	1	48
	Dagnum	13	17	1

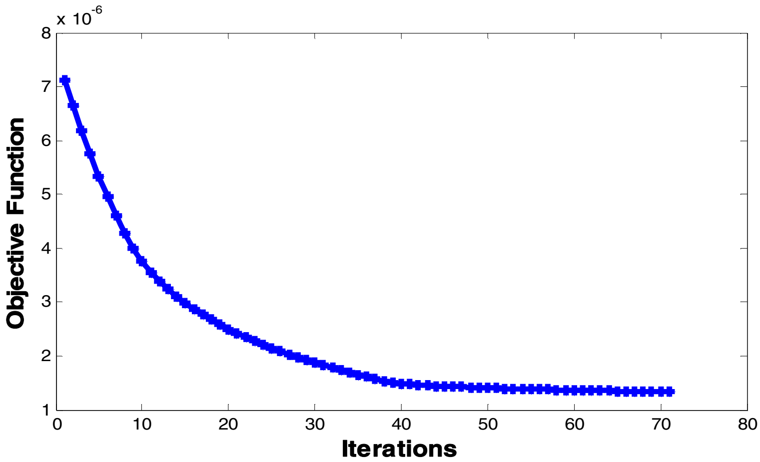

Fig. 3. Objective value as a function of the number of iterations for RBCOM applied to SY Corpus Development Problem (CDP)

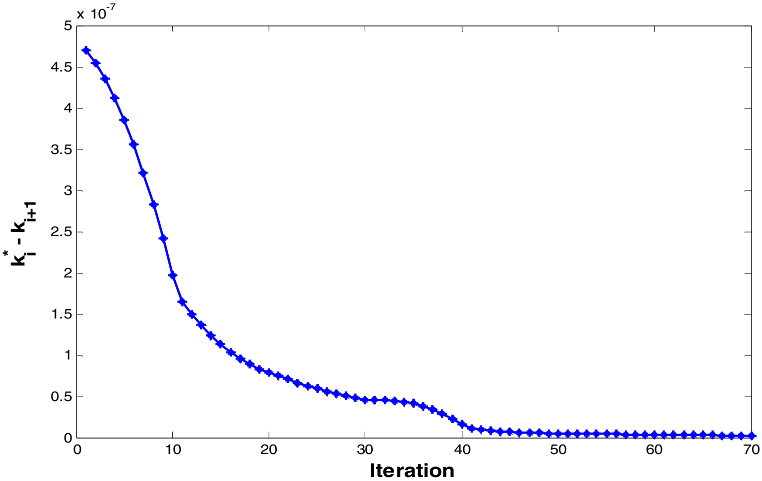

Fig. 4. Algorithm convergence as a function of the number of iterations for RBCOM applied to SY Corpus Development Problem (CDP)

A general trend of syllable occurrence and variants of distribution based on goodness of fit is presented in Table 2**.** The results of the expert model mining of allophone occurrence indicates that the Log Pearson 3, Gen. Extreme Value, Weibull, Burr, Lognormal, Pearson 5 (3P) and Log Logistic are the best performing models for Kolmogorov Smirnov, Andersen Darling and Chi-Squared tests.

3.2 Results of Random and RBCOM Generated Corpus

Figure 6 illustrates the syllable frequency for random and RBCOM corpus; the approximate syllable frequency range achieved for RBCOM prompt spans from 57 to 1200 while for the random scheme, the range spans from 0 to 6200. From the results, it is evident that random scheme cannot guarantee optimal allophone (syllable) coverage. Some syllables have a zero frequency of occurrence with very large frequency bandwidth. The RBCOM represents an improvement over the random scheme with low deviation of frequency distribution of syllable. This justifies the results presented in Figure 5 and Table 2.

A further analysis of Random and RBCOM corpus based on vague-linguistic terms are as presented in Graphs I-VI of Figure 7. The schemes were assessed based on three vague linguistic categories, namely low, middle and high range frequency distributions. Graphs I-VI of Figure 7 represents a performance profile of both Random and RBCOM corpus for low, middle and high range syllable frequency. For the low syllable frequency as depicted on Graphs I and II, the following outputs depicts the performance of both the Random and RBCOM corpus:16 syllables with zero frequency for random scheme, 35 syllables within 1-9 frequency range, and 109 syllables within 10-70 frequency range. The RBCOM schemes herein have their syllables frequency ranging from 155 to 162.

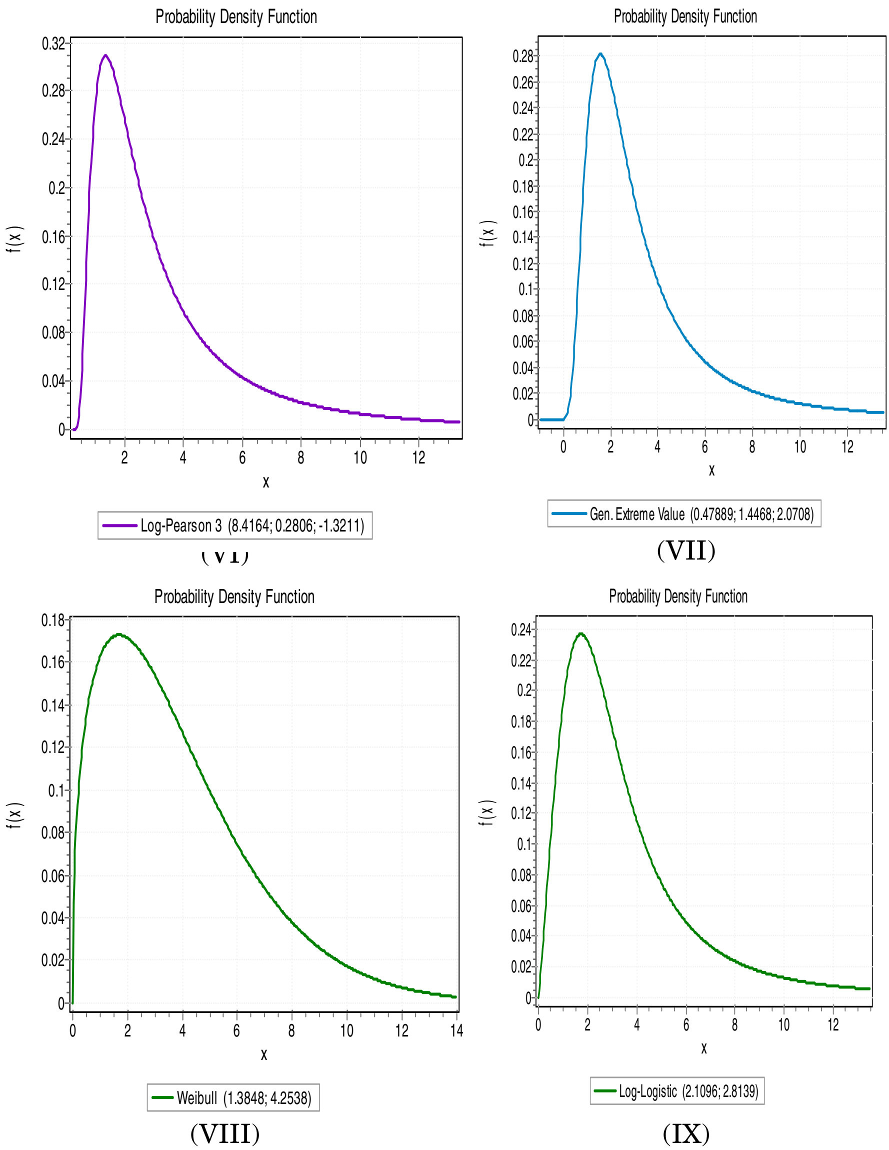

Fig. 5. SY syllable frequency trend for RBCOM prompt

Table 2. Expect Model Mining of Allophones for stable RBCOM corpus

S/No	Distribution	Kolmogorov Smirnov	Andersen Dar-ling	Chi-Squared
		Rank	Rank	Rank
TYPE I	Log Pearson 3	1	3	9
	Gen. Extreme Value	2	1	6
	Log. Logistics	6	8	1
TYPE II	Gen. Extreme Value	1	48	31
	Weibull	19	1	30
	Log Logistics	26	26	1
TYPE III	Weibull	1	10	3
	Burr	3	1	16
	Frechet	57	47	1
TYPE IV	Log Logistics	1	50	24
	Lognormal	8	1	7
	Pearson 5 (3P)	10	9	1

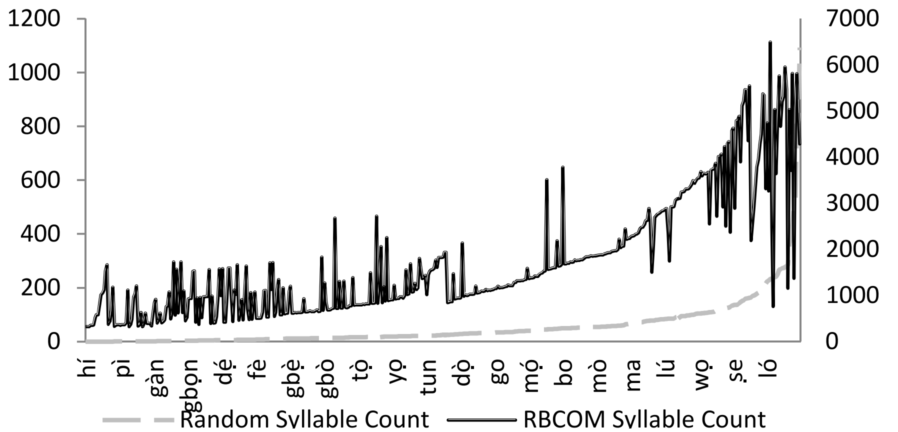

Fig. 6. Syllable Frequency Distribution

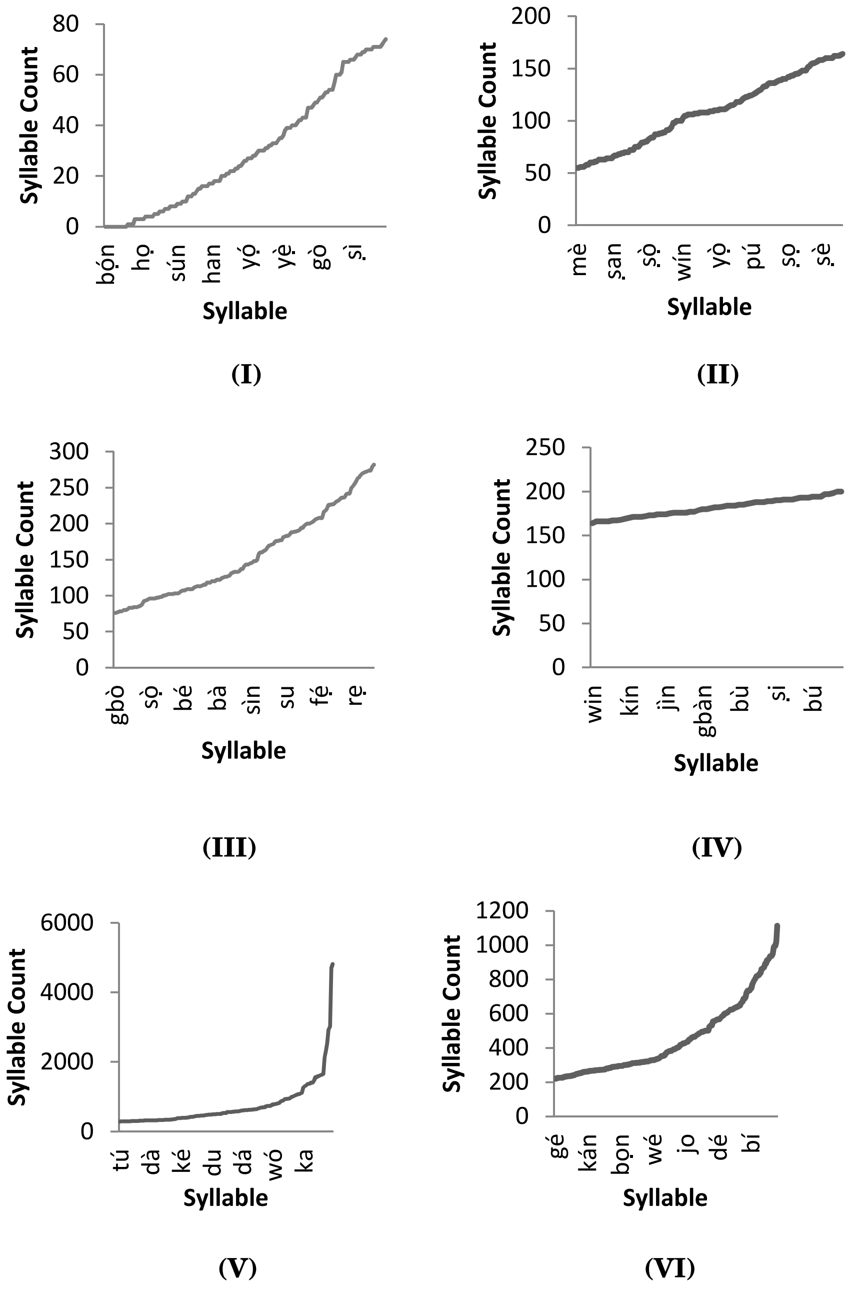

Fig. 7. Syllable Frequency against Syllable for Random and RBCOM Corpus Selection Strategy

Furthermore, in the case of the middle syllable frequency as obtained in Graphs III and IV of Figure 7, the performance of the Random and RBCOM corpus is premised on the following outcomes: availability of 120 syllables for the random scheme, frequency range of syllable is between 76 and 270 and RBCOM schemes have 61 syllables with syllable frequency range from 164-200. Also, Graphs V and VI of Figure 7 depicts the high syllable frequency which corresponds to the performance of both the Random and RBCOM corpus. The corresponding associated outputs includes: availability of 101 syllables for the random scheme, extension of the syllable frequency range from 282 to 6327 and availability of 241 syllables for the RBCOM schemes with syllable frequency range of 220-1113. These results have further stressed the initial assertion that corpus generated using the Random scheme are not phonetically balanced and also demonstrates the robustness of the RBCOM prompt selection strategy, in particular, for the moderate size corpus development of a resource-scarce language.

4 Conclusion

This paper has presented a new methodology for development of ASR corpus. The proposed technique which is premised on the use of Rule Based Corpus Optimization Model (RBCOM) has shown a higher prospect of results with respect to effectiveness. On comparison with the conventional randomized approach, the RBCOM demonstrated some level of superiority especially as displayed in the skewness of the distributed experimental data. While the random approach provided a skewness of less than -1, the RBCOM technique gave a skew distribution ranging from -1 to +1.

Furthermore, in classifying the allophone frequencies based on vague linguistic terms, the proposed RBCOM showed a better frequency range for each considered category of data unlike the random approach that generated a wider range for each class of allophones. For a future work, an allophone threshold for each genre of corpora is proposed.

References

1. Abushariah, M.A.A.M., Ainon, R.N., Zainuddin, R., Alqudah, A.A.M., Elshafei Ahmed, M., Khalifa, O.O.: Modern standard Arabic speech corpus for implementing and evaluating automatic continuous speech recognition systems. Journal of the Franklin Institute 349(7), 2215–2242 (2012)
2. Odéjobí, O.À.: A Quantitative Model of Yoruba Speech Intonation Using Stem-ML. INFOCOMP Journal of Computer Science 6(3), 47–55 (2007)
3. Adegbola, T., Owolabi, K., Odejobi, T.: Localising for Yorùbá: Experience, challenges and future direction. In: Proceedings of Conference on Human Language Technology for Development, pp. 7–10 (2011)
4. Àkànbí, L.A., Odéjobí, O.À.: Automatic recognition of oral vowels in tone language: Experiments with fuzzy logic and neural net-work models. Appl. Soft Comput. 11, 1467–1480 (2011)
5. Aibinu, A.M., Salami, M.J.E., Najeeb, A.R., Azeez, J.F., Rajin, S.M.A.K.: Evaluating the effect of voice activity detection in isolated Yorùbá word recognition system. In: 2011 4th International Conference on Mechatronics (ICOM), pp. 1–5. IEEE (May 2011)

6. Chomphan, S., Kobayashi, T.: Implementation and evaluation of an HMM-based Thai speech synthesis system. In: Proc. Interspeech, pp. 2849–2852 (August 2007)
7. Hoogeveen, D., Pauw, D.: CorpusCollie: a web corpus mining tool for resource-scarce languages (2011)
8. Cucu, H., Buzo, A., Burileanu, C.: ASR for low-resourced languages: Building a phonetically balanced Romanian speech corpus. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2060–2064. IEEE (August 2012)
9. Lecouteux, B., Linares, G.: Using prompts to produce quality corpus for training automatic speech recognition systems. In: MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference, pp. 841–846 (2008)
10. Nakamura, A., Matsunaga, S., Shimizu, T., Tonomura, M., Sagisaka, Y.: Japanese speech databases for robust speech recognition. In: Proceedings of the Fourth International Conference on Spoken Language, ICSLP 1996, vol. 4, pp. 2199–2202. IEEE (October 1996)
11. Metze, F., Barnard, E., Davel, M., Van Heerden, C., Anguera, X., Gravier, G., Rajput. N.: The Spoken Web Search Task. In: MediaEval 2012 Workshop, Pisa, Italy, October 4-5 (2012)
12. Lee, T., Lo, W.K., Ching, P.C., Meng, H.: Spoken language resources for Cantonese speech processing. Speech Communication 36(3), 327–342 (2002)
13. Abate, S.T., Menzel, W.: Automatic Speech Recognition for an Under-Resourced Language – Amharic. In: Proceedings of INTERSPEECH, pp. 1541–1544 (2007)
14. Raza, A.A., Hussain, S., Sarfraz, H., Ullah, I., Sarfraz, Z.: Design and development of phonetically rich Urdu speech corpus. In: 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, pp. 38–43. IEEE (August 2009)
15. Wu, T., Yang, Y., Wu, Z., Li, D.: Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In: IEEE Odyssey 2006 on Speaker and Language Recognition Workshop, pp. 1–5. IEEE (June 2006)