Greedy Algorithms to Optimize a Sentence Set Near-Uniformly Distributed on Syllable Units and Punctuation Marks

Bagus Nugroho Budi Nurtomo
School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu
Bandung, West Java, Indonesia 40257

Suyanto
School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu
Bandung, West Java, Indonesia 40257

***Abstract*—An optimum sentence set that near-uniformly distributed on syllable units and punctuation marks is important to develop a syllable-based automatic speech recognition (ASR). It is usually extracted from a mother set of millions of unique sentences using Modified Least-to-Most (LTM) Greedy algorithm. The Modified LTM Greedy is capable of minimizing the number of syllables but ignores distributing their frequencies. Hence, two schemes are proposed to minimize the number of syllables as well as to distribute their frequencies near-uniformly. Testing on a mother set of 10 million Indonesian sentences shows that both schemes perform better than the Modified LTM Greedy for two syllable units: monosyllables and bisyllables.**

***Keywords*—read-speech corpus; optimum sentence set; syllable; punctuation marks; Modified Least-to-Most Greedy**

I. Introduction

Since the begining 2000, some researchers show that the context-dependent syllable-based ASR systems perform better than the context-independent phone-based ones, as described in [1], [2], and[3]. Today, the promising state-of-the-art ASR called sequence-to-sequence attention-based model is also designed using a syllable-based model [4]. However, the syllable-based ASR needs a much larger read-speech corpus for the training process [5]. Therefore, developing such speech corpus is a challenging issue.

The speech corpus is commonly recorded on a minimum sentence set near-uniformly distributed on both syllable units and punctuation marks for thousands of speakers varying on gender, age, and dialect [6], [7], [8], [9], [10]. Punctuation marks in a sentence affect how it is being interpreted, mostly by differing intonation [11] and [12]. The speakers may use different intonation to make their intentions clear. A sentence "It's me." is a monotone statement, while "it's me?" gives a higher tone for the syllable 'me?'. Hence, a syllable-based ASR needs a read-speech corpus developed using a minimum sentence set balanced on syllables and punctuation marks [13] and [14].

Commonly methods used to extract a minimum sentence set from a mother set are greedy-based algorithms, such as the Least-to-Most (LTM) Greedy Algorithm [15]. This algorithm is then slightly improved to be the Modified LTM Greedy which is capable of extracting a minimum sentence set in quite fast execution time [16]. But, the Modified LTM Greedy only concentrates on minimizing the number of phonetic units but ignores balancing their frequencies.

In this paper, the Modified LTM Greedy is adapted to extract a minimum sentence set from a mother set of around 10 million sentences based on their syllable. Two additional schemes are proposed to make the Modified LTM Greedy capable of extracting a minimum sentence set, near-uniformly balanced on both syllables and punctuation marks, to be used to develop a state-of-the-art syllable-based ASR. Both additional schemes are carefully designed to minimize the number of syllables as well as to balance their frequencies.

II. Greedy Algorithms

The Modified LTM Greedy algorithm described in [16] performs well to extract a phonetically-rich sentence set. Unfortunately, it just focuses on minimizing the number of phonetic units but ignores balancing their frequencies. Hence, in this paper two additional schemes are proposed to improve the performance of the algorithm in minimizing the number of syllables as well as balancing their frequencies.

A. Modified LTM Greedy Algorithm

The Modified LTM Greedy algorithm produces a sentence set from a mother set by taking the best sentences based on a scoring formula. The pseudocode adapted from [16], with an adjusment to handle syllables instead of phonemes, is described as follows:

1) Let A = mother set, U = all to-be-covered syllables, B = empty set;
2) From U take all syllables with the lowest frequency and put them in U_{sub};
3) From A select all sentences containing at least one syllable in U_{sub} and put them in A_{sub};
4) Compute the score of each sentence in A_{sub} using a formula

$$S_i = \frac{N_i}{T_i}, \tag{1}$$

where S_i is the score for the *i*th sentence, N_i is the number of to-be-covered syllables in the *i*th sentence, and T_i is the number of all syllables in the *i*th sentence;

5) Choose a sentence with the best score and put it in B and remove all syllables contained in the sentence from both U and U_{sub};
6) Repeat step 3 to 5 until U_{sub} is empty;
7) Repeat step 2 to 6 until U is empty.

The pseudocode can be explained in a simple way using some illustrations in Fig. 1 to 5. In these illustrations, the mother set (A) contains only five sentences, as listed in Table I, to make any step in the pseudocode clear.

TABLE I. EXAMPLE MOTHER SET OF FIVE SENTENCES

Number	Sentence in *Indonesian* and (English)
1	*Belajar lagi di rumah* (Study again in home)
2	*Dia belajar video lagi* (He learns video again)
3	*Dia menonton di rumah belajar* (He is watching in the learning house)
4	*Lagi-lagi dia menonton di rumah* (Again he is watching at home)
5	*Menonton video di rumah* (Watching video at home)

In step 1, the Indonesian syllabification model described in [17] is used to generate all syllables contained in each sentence as well as a list of to-be-covered syllables, which contains 14 unique syllables, with their frequencies (U). The minimum set B is empty. Next, in step 2, all syllables with the lowest frequency in U are selected and moved into U_{sub}. In step 3, all sentences containing at least one syllable in U_{sub} are then selected and moved into A_{sub}. Then, in step 4, the score of each sentence in A_{sub} is calculated using the formula in Eq. 1. The second sentence, with 9 out of 10 to-be-covered syllables, has a score of 0.9. Meanwhile, the fifth sentence, with 9 out of 9 to-be-covered syllables, has a higher score of 1.0. Finally, in step 5, the fifth sentence with the best score of 1.0 is chosen, saved into B, and all syllables contained in this sentence are removed from both U and U_{sub}. These steps are repeated until both U_{sub} and U are empty. When both stoping criteria are reached the algorithm produces a minimum set of two sentences, i.e. the fifth and the second sentences, that consits of all 14 unique syllables to-be-covered.

A		
i	**Sentence**	**Syllables contained in the sentence**
1	Belajar lagi di rumah	be la jar la gi di ru mah
2	Dia belajar video lagi	di a be la jar vi de o la gi
3	Dia menonton di rumah belajar	di a me non ton di ru mah be la jar
4	Lagi-lagi dia menonton di rumah	la gi la gi dia me non ton di ru mah
5	Menonton video di rumah	me non ton vi de o di ru mah

U	
Syllable	**Frequency**
de	**2**
o	**2**
vi	**2**
a	3
be	3
jar	3
me	3
non	3
ton	3
gi	4
mah	4
ru	4
di	7
la	7

B	
Number	**Sentence**

Fig. 1. Step 1 of the Modified LTM Greedy algorithm: A = mother set, U = all to-be-covered syllables, B = empty set

U_{sub}	
Syllable	**Frequency**
de	2
o	2
vi	2

Fig. 2. Step 2 of the Modified LTM Greedy algorithm: take all syllables with the lowest frequency, i.e. 2, and put them in U_{sub}

A_{sub}		
i	**Sentence**	**Score**
2	Dia belajar video lagi	9/10 = 0.9
5	Menonton video di rumah	**9/9 = 1.0**

Fig. 3. Step 3 and 4 of the Modified LTM Greedy algorithm: select all sentences containing at least one syllable in U_{sub} and put them in A_{sub}, then compute the score of each sentence in A_{sub} using the formula in Eq. 1

B	
i	**Sentence**
5	Menonton video di rumah

U	
Syllable	**Frequency**
a	3
be	3
jar	3
gi	4
la	7

U_{sub}	
Syllable	**Frequency**

Fig. 4. Step 5 of the Modified LTM Greedy algorithm: choose a sentence with the best score, i.e. 1.0, and put it in B and remove all syllables contained in it from both U and U_{sub}

B	
i	**Sentence**
5	Menonton video di rumah
2	Dia belajar video lagi

U	
Syllable	**Frequency**

U_{sub}	
Syllable	**Frequency**

Fig. 5. Last steps of the Modified LTM Greedy algorithm, when both U_{sub} and U are empty, produce a minimum set of two sentences

B. Semi LTM Greedy 1

In the first proposed scheme, the Modified LTM Greedy is revised by replacing the step 5 with four new steps below:

1) Let K be a real number in the interval $(0, 1)$;
2) From A_{sub} select the top-score sentences, which have scores $\geq$ (the best score $\times(1-K)$), and put them in a new set D;
3) From D choose a sentence with the maximum number of to-be-covered syllables and remove all syllables contained in the sentence from both U and U_{sub};
4) Clear D.

This proposed scheme can be explained using an illustration in Fig. 6. In this illustration, let K = 0.05. From the mother set (A), which is sorted by the score calculated using the formula in Eq. 1, select the top-score sentences and put them into a new set D. Next, from D choose a sentence with the maximum number of to-be-covered syllables, i.e. 24, instead

of the highest score. This scheme is designed to handle the possibility of the Modified LTM Greedy algorithm in taking the local optimum when looking for the best sentence. It will produce a larger sentence set B.

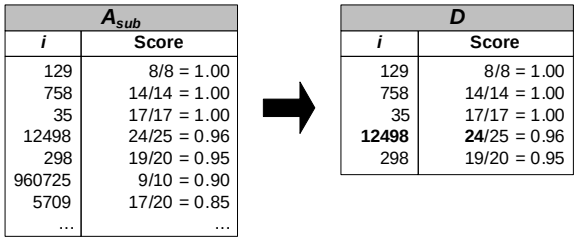

Fig. 6. Semi LTM Greedy 1: select the top score sentences in the mother set (A) and then choose a sentence with the maximum number of to-be-covered syllables

C. Semi LTM Greedy 2

In the second proposed scheme, the Modified LTM Greedy is updated by replacing the step 5 with four new steps below:

1) Let K be a real number in the interval $(0, 1)$;
2) Select the top-score sentences, which have scores $\geq$ (the best score $\times(1 - K)$), and put them in a new set D;
3) From D, choose a sentence with the lowest new score calculated using a formula:

$$S_i = \sum f, \tag{2}$$

where f is the frequencies of all have-been-covered syllables in the minimum set *B* and remove all syllables contained in the sentence from both U and U_{sub};
4) Clear D.

This scheme is proposed to overcome the weakness of the Modified LTM Greedy algorithm in balancing frequencies of the syllables. By taking sentences with the lowest frequencies of syllables have been covered in the minimum set *B*, the duplication of syllables should be reduced.

III. Experimental Setup

In this research, a mother set containing 10,000,034 sentences is collected by crawling some newspaper websites. Two dictionaries (phonemic and syllabic-based) of 80K unique words are developed using the Indonesian grapheme-to-phoneme conversion system described in [18] and the Indonesian syllabification system described in [17] respectively. Converting the mother set of 10 M sentences using both dictionaries produces 121,860,535 monosyllables (6,804 unique monosyllables) and 132,445,220 bisyllables (308,710 unique bisyllables).

Using the mother set, some experiments are performed based on two scenarios:

1) Scenario 1: The Modified LTM Greedy. In this scenario, the mother set is extracted using the Modified LTM Greedy for both monosyllable and bisyllable.
2) Scenario 2: The Semi LTM Greedy. In this scenario, the mother set is extracted using the Semi LTM Greedy 1 and the Semi LTM Greedy 2 with K = 0.05, 0.1, 0.2 and 0.33 for both monosyllable and bisyllable. The extracted minimum sentence sets balanced on syllables and punctuation marks are compared to those resulted by the Modified LTM Greedy.

IV. Result and Discussion

Two scenarios described in the experimental setup are tested for both monosyllables and bisyllables to compare their performances. The experiments are conducted using a single processor i5 with 4 GB RAM. The total run time per experiments for the monosyllables is 4 hours while for the bisyllables is 9 hours.

A. Monosyllable

Extraction of the mother set of 10 M sentences using the Modified LTM Greedy produces a sentence set of 6,804 unique monosyllables in 4,056 sentences with the total number of monosyllables is 31,575. The average frequency of syllable $\bar{f} = 4.64$ with the standard deviation $\sigma = 30.91$. Next, extraction of the mother set using the Semi LTM Greedy 1 and the Semi LTM Greedy 2 produce the results illustrated in Table II and Fig. 7.

TABLE II. Extraction of the Mother Set for Monosyllable

Exp.	Method	Tot. Syll.	Tot. Sent.	$\bar{f}$	σ
1	Modified LTM Greedy	31,575	4,056	4.64	30.91
2	Semi LTM 1, $K = 0.05$	31,754	4,030	4.66	31.15
3	Semi LTM 1, $K = 0.10$	31,905	3,985	4.86	32.66
4	Semi LTM 1, $K = 0.20$	33,115	3,950	4.50	31.40
5	Semi LTM 1, $K = 0.33$	34,688	3,956	5.09	34.62
6	Semi LTM 2, $K = 0.05$	31,560	4,087	4.63	29.68
7	Semi LTM 2, $K = 0.10$	31,666	4,160	4.65	29.48
8	Semi LTM 2, $K = 0.20$	32,272	4,277	4.74	28.64
9	Semi LTM 2, $K = 0.33$	33,537	4,471	4.92	28.85

TABLE III. Extraction of the Mother Set for Bisyllable

Exp.	Method	Tot. Syll.	Tot. Sent.	$\bar{f}$	σ
1	Modified LTM Greedy	2,453,766	202,157	7.94	83.09
2	Semi LTM 1, $K = 0.05$	2,455,142	201,877	7.95	83.09
3	Semi LTM 1, $K = 0.10$	2,451,017	201,609	7.93	83.00
4	Semi LTM 1, $K = 0.20$	2,466,306	201,840	7.98	83.42
5	Semi LTM 1, $K = 0.33$	2,477,568	201,962	8.02	83.81
6	Semi LTM 2, $K = 0.05$	2,456,045	202,586	7.95	81.64
7	Semi LTM 2, $K = 0.10$	2,460,139	202,786	7.96	81.51
8	Semi LTM 2, $K = 0.20$	2,468,850	203,159	7.99	81.43
9	Semi LTM 2, $K = 0.33$	2,471,132	203,387	8.00	81.15

Table II shows that the Semi LTM Greedy 1 is successful in reducing the total number of sentences, but it increases the total number of syllables as the value of K does. This is probably the case where the algorithm does not really consider the redundancy of syllables when taking the best sentence resulting in a large number of syllables.

On the other hand, the Semi LTM Greedy 2 is capable of reducing the standard deviation of the result set relatively as the K increases, but with the number of sentences increases as the K does. The formula used in the algorithm considers the frequencies of have-been-covered syllables and then takes the sentence with the smallest total frequencies. This prefers to select shorter sentences and make the result set larger. Fig. 7 shows that the Semi LTM Greedy 2 manages to lower the number of occurrences of more dominant syllables.

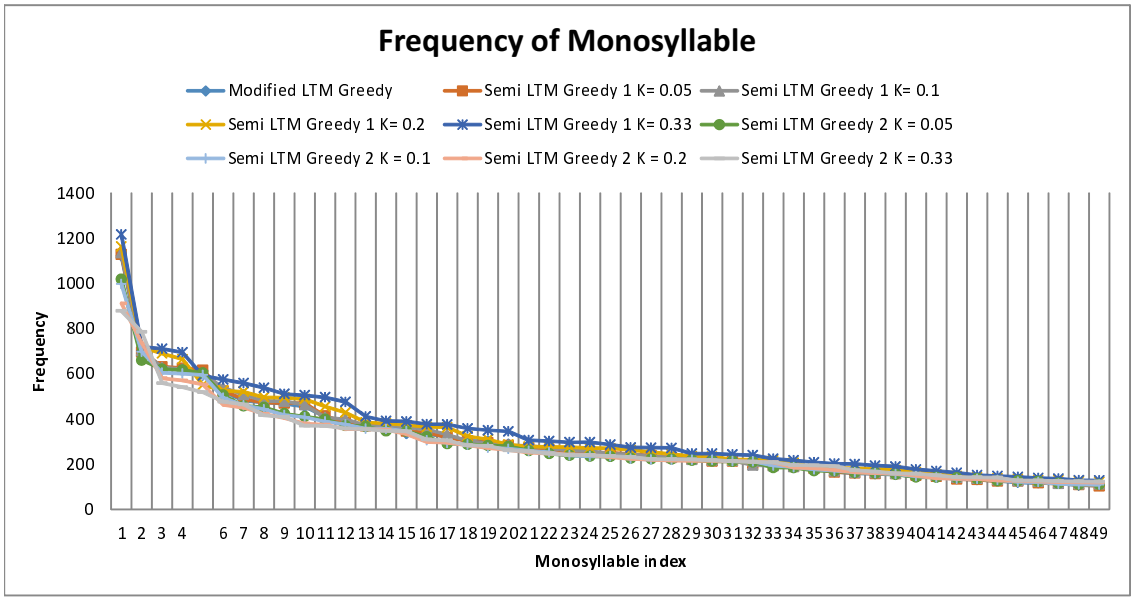

Fig. 7. Frequency of Monosyllable

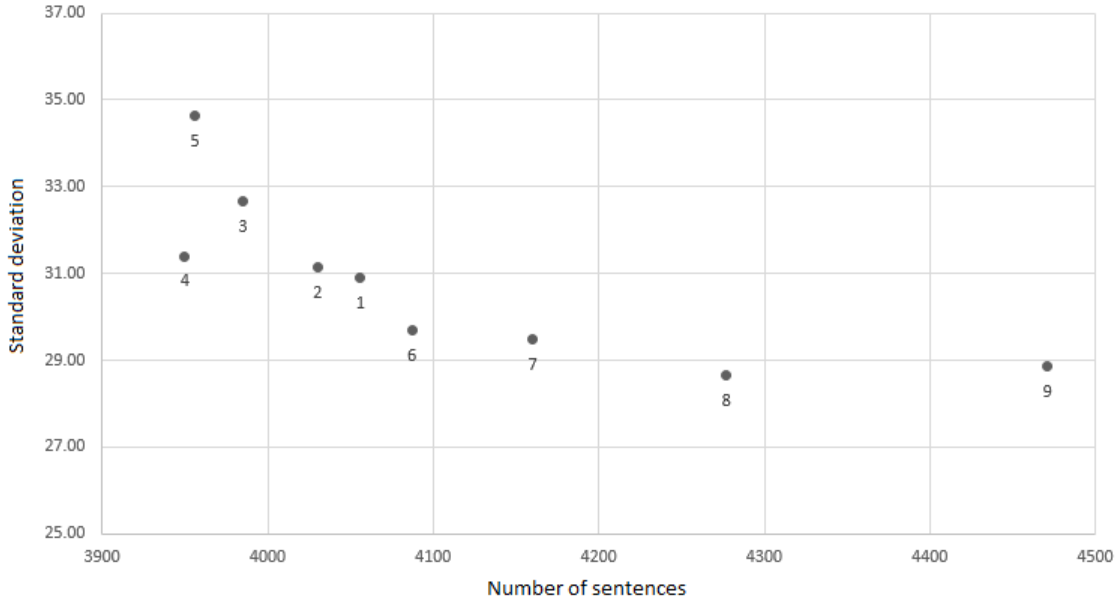

Fig. 8. Visualization of monosyllabic Pareto

A simple Pareto optimization in Fig. 8 shows that: a) Experiment 4 dominates Experiment 5 and 3; b) Experiment 8 dominates result 9; and c) Experiment 1, 2, 4, 6, 7 and 8 do not dominate each other. Hence, it can be concluded that the experiments 1, 2, 4, 6, 7 and 8 are the Pareto optimal set those should be able to be used as the train sets for the syllable-based ASR. The set from Experiment 6 should be used if the train set needs both low standard deviation and number of sentences. Experiment 4 has the smallest number of sentence set and best suited if the system demands as such while the result of Experiment 8 if requires as low standard deviation as possible.

B. Bisyllable

Using the mother set for bisyllables, the Modified LTM Greedy extracts a sentence set of 202,157 unique sentences with 308,710 unique bisyllables. The average frequency of bisyllable $\bar{f} = 8.94$ with the standard deviation $\sigma = 83.09$. Next, extraction of the mother set using the Semi LTM Greedy 1 and the Semi LTM Greedy 2 produces the results illustrated in Table III and Fig. 9.

Table III shows that the Semi LTM Greedy 1 manages to reduce the number of sentences and standard deviation using $K = 0.1$. The scenarios of the Semi LTM Greedy 2 show that it manages to reduce the standard deviation quite well, with the

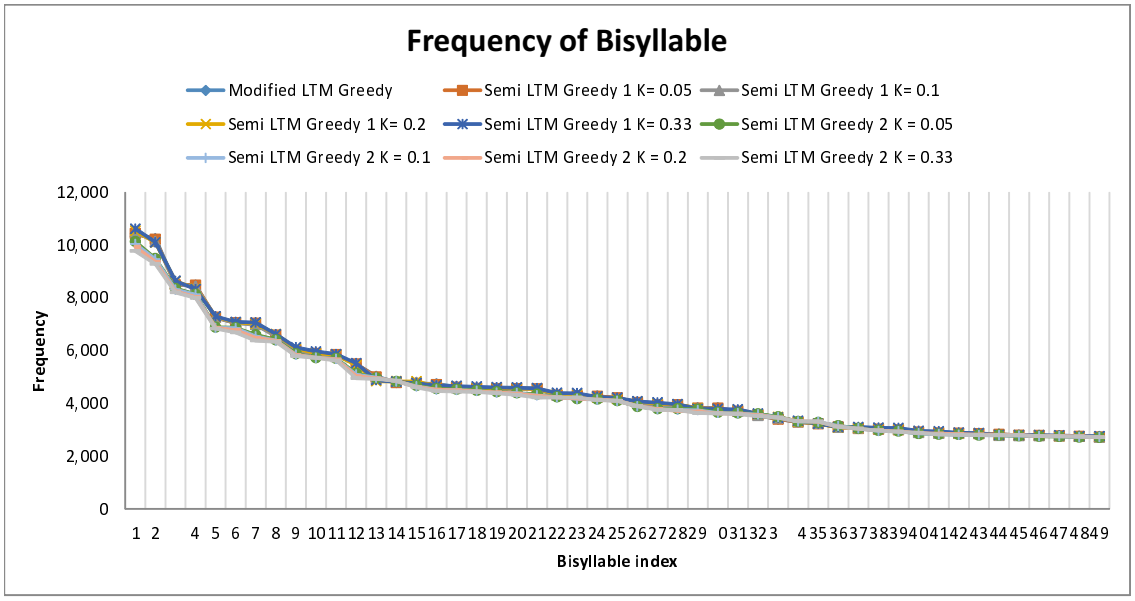

Fig. 9. Frequency of Bisyllable

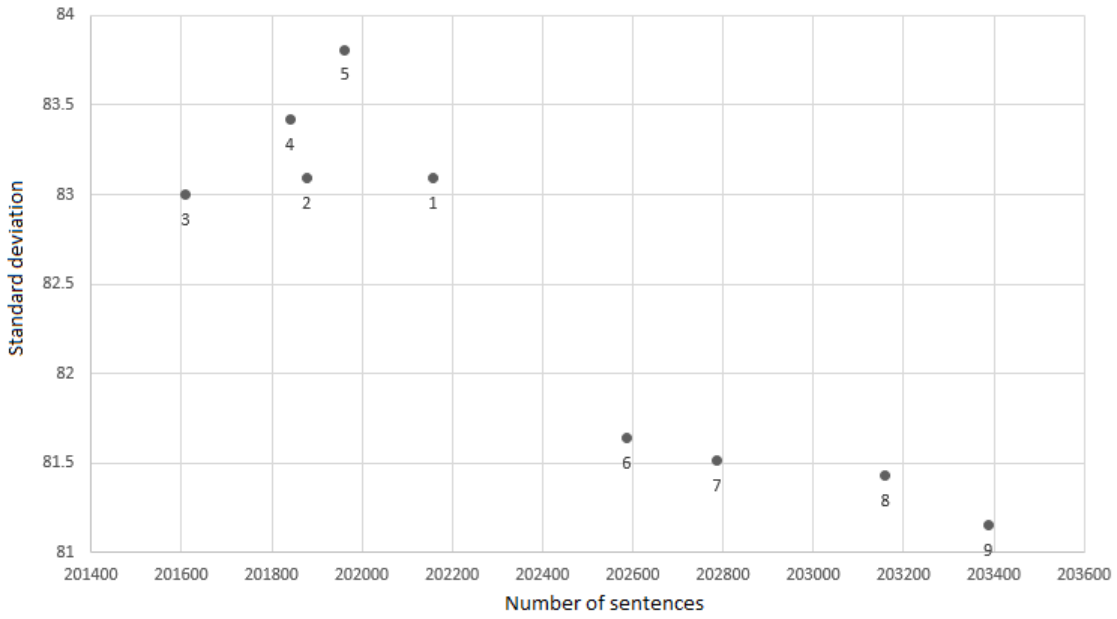

Fig. 10. Visualization of Bisyllable Pareto

scenario using $K = 0.05$ in particular. Increasing K reduces the standard deviation, but increases the number of sentences. Fig. 9 also shows that the Semi LTM Greedy 2 manages to reduce the frequencies of more dominant bisyllables, which should produce lower standard deviation.

A simple Pareto optimization using Fig. 10 shows that: a) Experiment 3 dominated experiments 1, 2, 4, and 5; and b) Experiments 3, 6, 7, 8, and 9 do not dominate each other. Thus, it can be concluded that the experiments 3, 6, 7, 8, and 9 are the optimum Pareto set those should be able to be used as a train set for a syllable-based ASR system. The sentence set from Experiment 6 should be used if the system needs a relatively low standard deviation and total sentences. Experiment 3 produces the set best suited for any system requiring as few sentences as possible, while Experiment 9 if least standard deviation.

V. Conclusion

The Semi LTM Greedy 1 algorithm is capable of reducing the number of sentences in the extracted sentence set, but the Semi LTM Greedy 2 manages to reduce standard deviation significantly. The Semi LTM Greedy 1 reduces more sentences as the K increases. The Semi LTM Greedy 2 reduces more standard deviation as the K increases. A simple Pareto opti-

mization can be used to produce the best sentence set for the designed syllable-based ASR.

ACKNOWLEDGMENT

The authors would like to thank Telkom University and colleagues for providing the mother set of 10 million sentences used in this research.

REFERENCES

[1] A. Ganapathiraju and J. Hamaker, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, 2001.

[2] R. Janakiraman, J. C. Kumar, and H. A. Murthy, "Robust syllable segmentation and its application to syllable-centric continuous speech recognition," in *National Conference on Communications (NCC)*. Chennai, India: Joint Telematics Group of IITs & IISc, jan 2010, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5430189

[3] K. Proen, "Designing Syllable Models for an HMM Based," in *International Conference on Speech and Computer*, vol. 1, 2016, pp. 216–223.

[4] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese," in *Interspeech*, 2018.

[5] ——, "A Comparison of Modeling Units in Sequence-to-Sequence Speech Recognition with the Transformer on Mandarin Chinese," *CoRR*, pp. 1–5, 2018.

[6] C. Kurian, "Development of Speech corpora for different Speech Recognition tasks in Malayalam language," in *International Conference on Natural Language Processing*, no. December, 2015, pp. 229–236.

[7] M. Pinnis, A. Salimbajevs, and I. Auzia, "Designing a Speech Corpus for the Development and Evaluation of Dictation Systems in Latvian," in *The Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016, pp. 775–780.

[8] D. Arnold, F. Tomaschek, K. Sering, F. Lopez, and R. H. Baayen, "Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit," *PLOS ONE*, vol. 12, no. 4, pp. 1–16, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0174623

[9] D. Koržinek, K. Marasek, Ł. Brocki, and K. Wołk, "Polish Read Speech Corpus for Speech Tools and Services," in *CLARIN*, 2017, pp. 54–62.

[10] H. Abera and S. H/Mariam, "Design of a Tigrinya Language Speech Corpus for Speech Recognition," in *Workshop on Linguistic Resources for Natural Language Processing*, 2018, pp. 78–82.

[11] J. Kolár and L. Lamel, "On Development of Consistently Punctuated Speech Corpora," in *INTERSPEECH*, 2011, pp. 833–836.

[12] N. Moore, "What ' s the point ? The role of punctuation in realising information structure in written English," *Functional Linguistics*, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40554-016-0029-x

[13] F. Batista, D. Caseiro, N. Mamede, I. Trancoso, L. F. L. D. Sistemas, D. L. Falada, I. Id, and R. A. Redol, "Recovering Punctuation Marks for Automatic Speech Recognition," in *Interspeech*, 2007, pp. 2153–2156.

[14] S. M. Hosseini and H. Sameti, "Creating a corpus for automatic punctuation prediction in Persian texts," in *2017 Iranian Conference on Electrical Engineering (ICEE)*, 2017, pp. 1537–1542.

[15] J.-s. Zhang and S. Nakamura, "An Efficient Algorithm to Search For A Minimum Sentence Set For Collecting Speech Database," in *ICPhS*, 2003, pp. 3145–3148.

[16] Suyanto, "Modified Least-to-Most Greedy Algorithm to Search a Minimum Sentence Set," in *IEEE TENCON*, 2006.

[17] S. Suyanto, S. Hartati, A. Harjoko, and D. V. Compernolle, "Indonesian syllabification using a pseudo nearest neighbour rule and phonotactic knowledge," *Speech Communication*, vol. 85, pp. 109–118, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2016.10.009

[18] Suyanto and A. Harjoko, "Nearest neighbour-based Indonesian G2P conversion," *Telkomnika (Telecommunication, Computing, Electronics, and Control)*, vol. 12, no. 2, pp. 389–396, 2014.