Segmentation of Malay Syllables in Connected Digit Speech Using Statistical Approach

M-S Salam sah@utm.my
Faculty of Computer Science & Information System
University Technology Malaysia
Skudai, 81310 Johor Bahru, Malaysia

Dzulkifli Mohamad dzul@utm.my
Faculty of Computer Science & Information System
University Technology Malaysia
Skudai, 81310 Johor Bahru, Malaysia

S-H Salleh hussain@fke.utm.my
Faculty of Biomedical Engineering & Health Science
University Technology Malaysia
Skudai, 81310 Johor Bahru, Malaysia

Abstract

This study present segmentation of syllables in Malay connected digit speech. Segmentation was done in time domain signal using statistical approaches namely the Brandt's Generalized Likelihood Ratio (GLR) algorithm and Divergence algorithm. These approaches basically detect abrupt changes of energy signal in order to determine the segmentation points. Patterns used in this experiment are connected digits of 11 speakers spoken in read mode in lab environment and spontaneous mode in classroom environment. The aim of this experiment is to get close match between reference points and automatic segmentation points. Experiments were conducted to see the effect of number of the auto regressive model order *p* and sliding window length *L* in Brandt's algorithm and Divergence algorithm in giving better match of the segmentation points. This paper reports the finding of segmentation experiment using four criterions ie. the insertion, omissions, accuracy and segmentation match between the algorithms. The result shows that divergence algorithm performed only slightly better and has opposite effect of the testing parameter *p* and *L* compared to Brandt's GLR. Read mode in comparison to spontaneous mode has better match and less omission but less accuracy and more insertion.

Keywords: Speech Segmentation, Divergence Algorithm, Brandt's Algorithm

1. INTRODUCTION

Malay language is an agglutinative language. It is a language of derivative, which allows addition of prefix and suffix to the base word to form new word(s) [1]. Most of Malay words can be

considered as consist of combination of syllables where syllables can be comprised of a vowel, or a vowel with a consonant or a vowel with several consonants [2] Several studies and experiments show that syllable unit size is remarkably salient and may exhibit specific acoustic characteristic [3]. Being able to segment the syllables correctly will make recognition a much easy work. Previous experiment on isolated Malay digit syllables where the segmentation was done manually reach recognition up to 80% [4]. However, automatic syllable segmentation from connected digit is a taunting task as syllables signal in connected speech is highly complex with no fixed property and significant acoustic cues exists in between syllables.

Time domain segmentations with non-fixed overlapping segment window size proved to give a good segmentation result with less omission [5]. Among these non-fixed overlapping segment window size segmentation, two algorithms usually applied are the Brandt's GLR algorithm and the Divergence algorithm. Brandt's GLR algorithm and divergence algorithm detect segment points by identifying discontinuities of speech signal without any further knowledge upon the phonetic sequence [6]. In another words they are linguistically unconstrained and are therefore expected to make insertions and omissions. Nevertheless, an "ideal" Brandt's GLR which disregards omission and insertions yields better word segmentation accuracy compare to HMM in experiment done in [6]. On the other hand, experiment on segmentation of music found that divergence algorithm performed better than Brandt's algorithm [7]. Similar conclusion is yielded for experiment on word in continuous speech in [8].

With respect to the foregoing, this paper report works in syllable segmentation from a sequence of Malay connected digits speech signal using both Brandt's GLR and divergence algorithms with the objective to find the best match between automatic segmentation and reference segmentation points. The aim is to apply the points from automatic segmentation of syllables in recognition of connected digit. That work however, is beyond of the scope of this paper.

Four evaluation criterions are subjects of this paper interest which are the omissions, insertion, accuracy and match based on given time tolerance. The experiment conducted on different value of auto regressive model order p and sliding window length L is to analyze the effect of these parameters upon those criterions and speech utterance mode. This report is outline as follow. Next section describes human perception of word, and then the data used in this experiment which is Malay Connected Digits is explained in the next section. The following section after that is on approaches applied in the experiment. The result is reported in the following section with discussion and conclusion at end of the report.

2. HUMAN PERCEPTION OF WORD

Word pronunciation and perception are common task for human. In daily communication syllables in word are not pronounced clearly and at equal phase which lead to lack of acoustic information of the word. Human however can easily anticipate the incomplete information at perception level. This is so because in real time human communication, human does not listen speech utterances in complete but anticipate them by comparing some existing sound model in their brain [9][10]. Human perception is not only accurate but also rapid [11]. When a word is said, human will listen and able to segment chunk by chunk of acoustic information before making perception of the word. In most cases, a native speaker of a language would already know what is going to be said prior to end of the word uttered in that language.

As an example, the word *senaman* can be percept by human as stages at Table 1. When the sound /s/ is heard, native listener would already have in his brain a list of possible word starting with sound /s/. The list reduced as the acoustics information of the sound become clearer. In this straight forward example, the listener already percept the word at stage 5 before the word is fully uttered.

Stage	Sound heard	Optimization of the word
1	/s/	**s**atu, **s**aya, **s**emut, **s**enapang, **s**enak, **s**eniman, **s**enada, **s**enaman **s**ilat, **s**ikap, **s**opan, etc… (any words in memory that start with /s/)
2	/se/	**se**mut, **se**napang, **se**nak, **se**niman, **se**nada, **se**naman (only word start with /se/ remain)
3	/sen/	**sen**apang, **sen**ak, **sen**iman, **sen**ada, **sen**aman (reduced to fewer words)
4	/se/ +/na/	**sena**pang, **sena**da, **sena**man (fewer possible words)
5	/se/ + /nam/	**senam**an (already anticipate the word)
6	/se/+/nam/+ /a/	**senama**n
7	/se/+ /nam/ + /an/	**senaman**

TABLE 1: Stages in Human Perception of The Word *senaman*

Some researcher works on phoneme as the basic chunk of acoustic information deriving the word [12][13]. Phoneme based however, is too fragile as to have only very small interval. Thus is not suitable for integration of spectral and temporal dependencies [14]. Furthermore, phoneme segmentation is much more difficult compare to syllables due to the same reasons. This work emphasizes on syllables as the basic acoustic chunk for segmentation and perception as it is salient and may exhibit specific acoustic characteristic especially in Malay language.

3. THE SYLLABLES IN MALAY DIGIT

There are 20 syllables that consist in Malay digit from 0-9. The syllables are { ko\, song\, sa\ , tu\ ,du\, a\, ti\, ga\, em\, pat\, li\, ma\, e\, nam\, juh\, la\, pan\, sem\, bi\, lan\ }. Table 2 shows the combination of phonetic alphabets in the 20 syllables. In general there are combinations of four types of consonant alphabets which are plosive, nasal, fricative and lateral approximant consonants and two types of vowels which are front and back vowels.

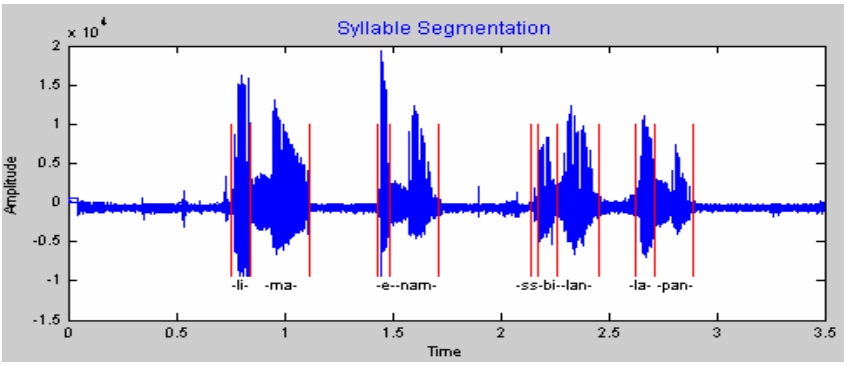

FIGURE 1: Connected Digit and its Syllable Manual Segmentation.

These syllables are visually and audibly distinguishable when pronounced in clear read mode. In spontaneous mode on the other hand, the cues are not as clear. Therefore, even for manual segmentation by human the task is not easy. Figure. 1 shows example of connected digit and its manual segmentation for digit ”lima-enam-sembilan-lapan” (5698).

Most of the syllables have two significant energy clusters where the start and end of the syllables is visually noticeable based on the abrupt changes of the energy. However, when pronounced connectedly and closely together, the correct segmentation point do not significantly visual and false abrupt changes exist in other places depends on the speaker's utterance style. The syllable's sound and the signal pattern visually are also influenced by the preceding and following syllables. This effect is known as co-articulation effect.

No.	Syllables	Descriptions
1	/a/ and /e/	A front vowel syllable
2	/bi/, /ga/ and /ti/	A plosive consonant with a front vowel
3	/ko/, /du/ and /tu/	A plosive consonant with a back vowel
4	/ma/	A nasal consonant with a front vowel
5	/pan/	A plosive consonant with a front vowel ending with a nasal consonant
6	/nam/	A nasal consonant with a front vowel ending with a nasal consonant
7	/sem/	A fricative consonant with a front vowel ending with a nasal consonant
8	/pat/	A plosive consonant with a front vowel ending with plosive
9	/la/	A lateral approximant consonant with a front vowel
10	/sa/	A fricative consonant with a front vowel
11	/em/	A front vowel with a nasal consonant
12	/li/	A lateral approximant consonant with a front vowel
13	/lan/	A lateral approximant consonant with a front vowel ending with a nasal consonant
14	/juh/	An lateral approximant consonant with a back vowel ending with a fricative consonant
15	/song/	A fricative consonant with a back vowel ending with a nasal and plosive consonants

TABLE 2: Phonetic Attributes of The 20 Syllables.

For read mode patterns, most of the pattern signal is quite clear as there are significant silence intervals in between words and even syllables for some cases. The only minimum noises are from nasal, mouth and lips ie. there is no background noise. However, for spontaneous mode patterns It is much complex as the position of the words does not necessary have silence interval and noises from back ground exist which leads to extra abrupt changes of the energy. It is observed that syllable with consonant like 'p','l' and 't' make short instance fluctuation in energy signal. These extra fluctuations of energy may lead to difficulties in obtaining the right point and increase number of insertion in segmentation.

4. THE STATISTICAL APPROACHES

Both Brandt's GLR algorithm and Divergence algorithm use statistical analysis in determining the segment points. The signal is assumed to be described by a string of homogeneous units, each of which is characterized by a statistical model of the form:

$$y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n \qquad \text{(i)}$$

where e_n is the excitation of the acoustic channel and is an uncorrelated zero mean Gaussian sequence with $\text{var}(e_n) = \sigma_n^2$ The model is parameterized by the vector Θ defined by:

$$\Theta^T = (\theta^T, \varphi^T)$$
$$\theta^T = (a_1, ..., a_p) \qquad \text{(ii)}$$

where φ is parameter vector which determines the sequence σ_n. These methods consist in performing on line a detection of changes in the parameter Θ starting from location of the previous detected. The algorithms are basically, (1). Detect when changes occurs. (2). Estimate the location of the changes. The two segmentation algorithms differ in the assumption of the excitation of the model and in the choice of the test statistics. Basically, a fixed and a growing window are used and then suitable distance estimation compares the two spectra.

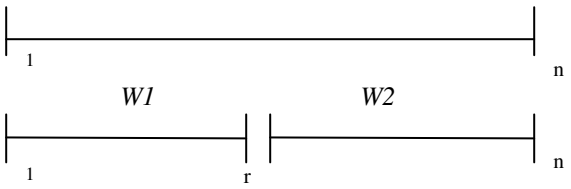

FIGURE 2: Location of the three windows in Brandt's GLR algorithm

4.1 Brandt's GLR algorithm

In Brandt's algorithm, the model assume

$$\sigma_n^2 = \sigma_n \quad \text{(i.e. } \varphi = \sigma \text{)} \qquad \text{(iii)}$$

The test is to monitor ($y_1,\ldots,y_n$), decide between the hypotheses:

- $H_o : \Theta = \Theta_0 \qquad for \quad 1 \le k \le n$
- $H_1 : \exists\, r \quad such\ that \quad \Theta = \Theta_1 \ for\ 1 \le k \le r$
  $and \ \ \Theta = \Theta_2 \quad for\ 1 \le k \le n$

There are 3 windows to manage as shown in Figure 2. The algorithm attempts to decide based on the likelihood between the two hypotheses, where the time instant r and the Θ_i's are replaced by their maximum likelihood estimates, so that the changes is detected if the distance

$$D_n = \max_r \max_{\Theta_1 \Theta_2} \max_{\Theta_0} \log\left(\frac{p\left[y_{1,\Lambda,} y_n \| H_1\right]}{p\left[y_{1,\Lambda,} y_n \| H_0\right]}\right) \ge \lambda \qquad \text{(iv)}$$

where λ is the threshold.

Then the estimate of the change $\hat{r}$ is the argument of the maximum in the relation (iv). The maximum likelihood estimates of the Θ's are given by the formulae:

$$\hat{\theta}(W_j) = \arg\min_\theta \sum_{k \in W_j} \left(Y_k - \sum_{i=1}^{p} a_i y_{k-i}\right)^2 \qquad \text{(v)}$$

$$\hat{\sigma}^2 = \min_\theta \frac{1}{card(W_j)} \sum_{k \in W_j} \left(Y_k - \sum_{i=1}^{p} a_i y_{k-i}\right)^2 \qquad \text{(vi)}$$

Where *W* denotes one of the three windows depicted in Figure 2. This finally yields the following formula for D_n.

$$D_n = \max_r D_n(r) \qquad \text{(vii)}$$

$$D_n = n \log \hat{\sigma}_0 - r \log \hat{\sigma}_1 - (n-r) \log \hat{\sigma}_2 \qquad \text{(viii)}$$

To avoid high computational cost in detection-estimation of the above formula, the different parameter Θ_0, Θ_1 and Θ_2 are identified by *Durbin-Levinson algorithm*.

4.2 Divergence Algorithm

The model set for divergence is similar as in Brandt's BLR algorithm. Equations (i), (ii) and (iii) are applied. In Divergence algorithm, the test is based on the monitoring of a suitable distance measure between the two models Θ_0 and Θ_1 located as shown in Figure3

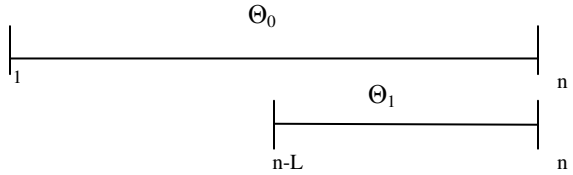

FIGURE 3: Location of The Two Models for The Divergence Algorithm

This distance is derived from the cross entropy between the conditional distribution of these two models. Consider

$\mathbf{y}_m^T = (y_{1,\mathrm{K}}, y_m)$ and denote by

$\vartheta_0(y_m \| \mathbf{y}_{m-1})$ *and* $\vartheta_1(y_m \| \mathbf{y}_{m-1})$

the two conditional densities corresponding to the models of Figure 3. Introduce the cross entropy between the two models, ϑ_0 and ϑ_1:

$$w_m = \int \vartheta_0(y \| \mathbf{y}_{m-1}) \log \frac{\vartheta_1(y \| \mathbf{y}_{m-1})}{\vartheta_0(y \| \mathbf{y}_{m-1})} dy - \log \frac{\vartheta_1(y \| \mathbf{y}_{m-1})}{\vartheta_0(y \| \mathbf{y}_{m-1})}$$

Which introduce the cumulative sum $W_n = \sum_{m=1}^{n} w_m$

It can be shown under hypothesis H_0: $\Theta = \Theta_0$ and under hypothesis H_1: $\Theta = \Theta_1$.
A change detection occur when the long term model disagree with the short term model in the sense of cumulative sum statistics. Detection is done by comparing the cumulative sum with threshold value as follow

$$\max_{1 \le r \le n} \hat{W}_r - \hat{W}_r > \lambda \quad \text{and}$$

$$W_n = \sum_{m=1}^{n} w_m + \delta$$

where δ is a bias value and λ is a threshold.

5. EXPERIMENTAL PROCEDURE

Our purpose is to see the affect of number of the auto regressive model order *p* and sliding window length *L* in Brandt's algorithm and Divergence algorithm to the four measurement criterions which are the omission, insertion match and accuracy. These criterions are evaluated by comparing the points gotten from automatic segmentation using the methods with a manual procedure. Hereon it will be known as referenced points. Points by manual procedure is considered the best as it used humanly ability consist of visual and audio intelligent that is through viewing the waveform pattern abrupt changes and verified through listening. Figure 4 shows an example of automatic and manual segmented pattern and the insertions and omission points. The graph is plotted using SFSWIN ver 1.5 from University College of London.

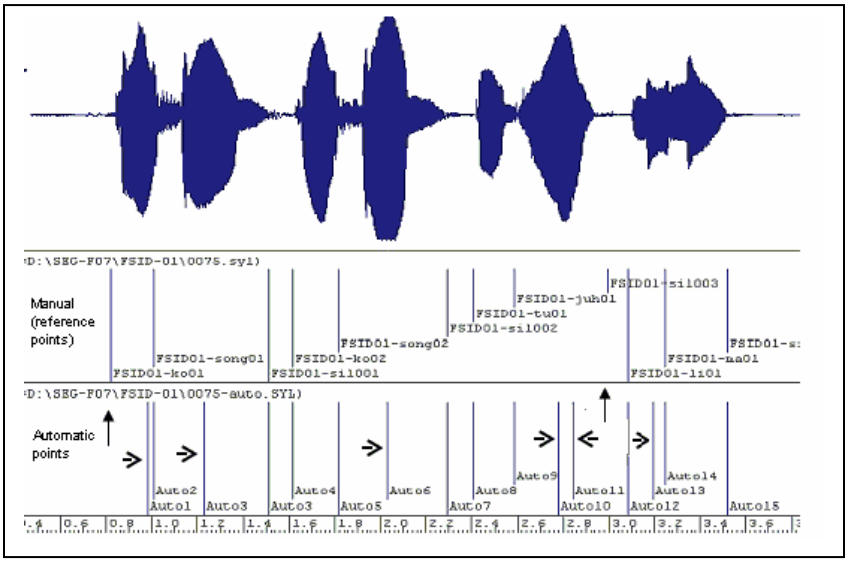

FIGURE 4: Example of comparison between reference (above) and automatic (below) segmentation points and its corresponding omission and insertion points. ↑ indicates omission points while → and ← shows insertion points.

The measurement criterions is defined as below adapted from [6],
Let $U = \{U_1, U_2, \ldots U_n\}$ and $V = \{V_1, V_2, \ldots V_p\}$ be the points in second of the segmentation marks obtained respectively by an automatic algorithm and by manual procedure which acts as the reference segmentation points. For each U_j, a correspondence is done with the reference segmentation by determining the time instant V_{kj} which is closest to U_j. A sequence $Vu = \{ V_{k1}, V_{k2}, \ldots V_{kn}\}$ is built in order to compare both segmentations.

Thus, omission is evaluated as points in *Vu* that is not in U_j and insertion is defined as extra points in U_j that is not in *Vu*. Match is calculated as number of similar points in U_j and *Vu* say, *m* divide by number of points in *V*, *p*. Thus, it can be defined as, match = $(m/p * 100)$ and accuracy = $((m/p+n) * 100)$ where accuracy will be influenced by number of insertion occurrences. Performance of the methods is evaluated better if has less omission and insertion and high match and accuracy. The value for threshold, λ for Divergence and Brandt were delivery set low as to be able to avoid omission and thus get better match however, it may lead to high insertion occurrences. Nevertheless, insertion is not our main concern in this work.

6. EXPERIMENTAL RESULT

The results of the experiments are presented in two sections. The first is on comparison between Brandt's and divergence algorithm and secondly on comparison between read mode and spontaneous mode. Comparisons are made in term of the performance criterion stated earlier in

the effect of the changes made on two experimental variables which are the auto regression model order *p* and sliding window length, *L*.

6.1 Comparison between Divergence and Brandt's GLR algorithms

The result shows that Brandt's has opposite effect to divergence algorithms in experiment on sliding window size *L* for all four criterion that are match, accuracy, omission and insertion. Incrementing the size *L* from 300 to 500 sample lead to better match, better accuracy and less omission but greater insertion for Brandt's algorithm. On the hand, for divergence algorithm the effect would be fewer matches, less accuracy, greater omission and lesser insertion.
Similar effects is observed in the experiment on the value for auto regression model order, *p*. Incrementing the value *p* increase accuracy and match, lessen omission but increase the insertion for Brandt's. In contrast for divergence, it leads to decreasing accuracy and match, increase the omission and decrease the insertion.

In general, there are only slightly different of better match observed in divergence algorithm compare to Brandt's in all experimental parameters. The increment or decrement probability different are very small around 0.005 to 0.20 percent for both algorithms. Figure 5 and Figure 6 show the graphs comparing the best match for both algorithms on spontaneous and read mode speech for experimental on *p* and *L*. The figures shows divergence algorithm perform better match with parameter *p*=2 and *L*=300 and Brandt with *p=2* at *L*=500.

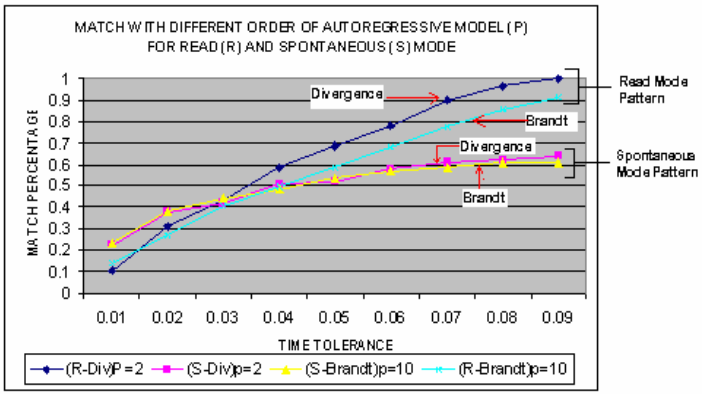

FIGURE 5: Match comparison between the best value for *p* for Divergence and Brandt on spontaneous and read mode patterns.

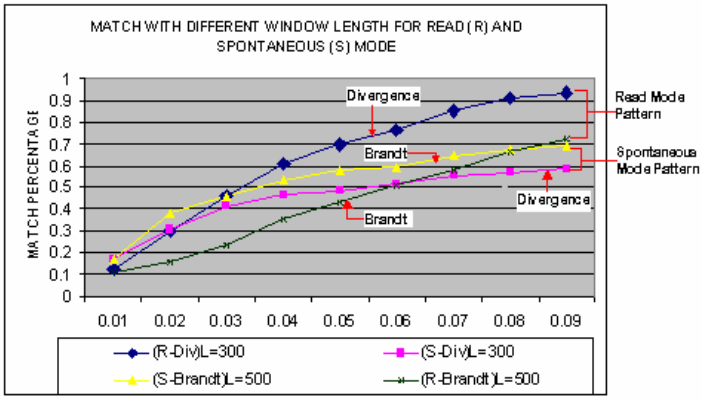

FIGURE 6: Match comparison between the best value for *L* for Divergence and Brandt on spontaneous and read mode patterns.

6.2 Comparison between Read Mode and Spontaneous Mode

It is expected that read mode segmentation would be easier and thus perform better than spontaneous mode. However, the experiment conducted observed that for certain criterion spontaneous mode has better performance than read mode. Number of insertion in spontaneous mode is less compare to read mode which lead to accuracy calculation for spontaneous mode better than read mode. On the other hand omission and match of spontaneous mode are not really good. The best match for spontaneous mode is 70% for L=500 and p=2 at time tolerance 0.09 second while for read mode it is 100% for p=2 and L=300 at time tolerance 0.09 using divergence algorithm. Similar to insertion with accuracy, omission goes less when match is high. The best accuracy for both modes is obtained using divergence algorithm is 44% for spontaneous mode and 42% for read mode. It can be noticed that accuracy criteria for read mode drop significantly compare to its match criteria due to accuracy calculation influenced by number of insertion occurrences. Figure 7, Figure 8 and Figure 9 show the best experimental result of accuracy, insertion and omission respectively for both read mode and spontaneous mode.

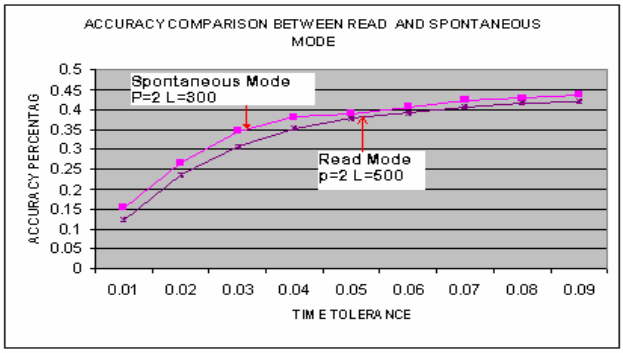

FIGURE 7: Accuracy comparison between spontaneous and read mode patterns.

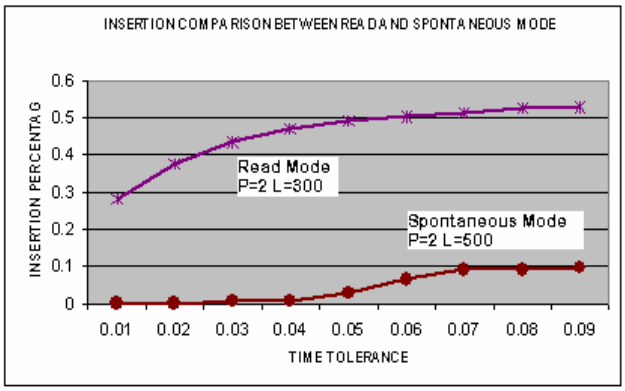

FIGURE 8: Insertion comparison between spontaneous and read mode patterns.

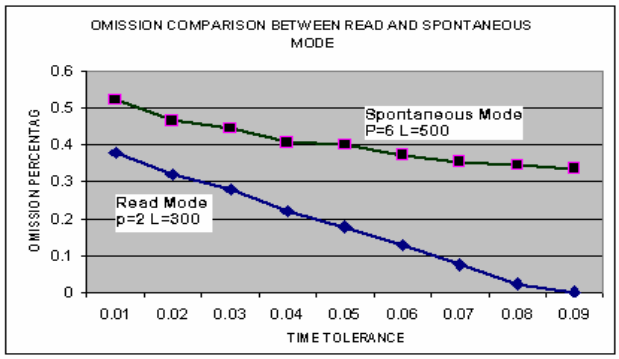

FIGURE 9: Omission comparison between spontaneous and read mode patterns.

7. CONCLUSION & FUTURE WORK

Segmentation is an inherently extremely a difficult problem [15]. The statistical approaches used in this experiment do not use any acoustic information in determining the segmentation point. Furthermore, the threshold value is set to low as not to miss any match in reference pattern. Thus, insertion is expected to occur. Nevertheless, the objective of getting no omission is achieved as segmentation match reach up to 100% using divergence algorithm.

Malay is an agglutinative language where words forming are combination of syllables or phoneme. Phoneme based modeling is too fragile as to have very small interval thus not suitable for integration of spectral and temporal dependency [14]. Whereas, syllable is able represent specific acoustic characteristic thus maybe most suitable as the basis in forming Malay words. Previous works in Malay isolated syllable recognition able to reach more than 80% recognition. However, no works has done to segment connected syllables in connected words. This work is the initial step to segment syllables as the basis for recognition of continuous Malay words. It is our future plan to develop an intelligent algorithm that can guess missing syllables with embedded language knowledge in the system. Our test using simulated data on the prototype system shows a promising result [16]. However, the result from this works indicates that insertion may become a drawback for real speech data. On going work is done to eliminate the insertion using Neural Network.

8. REFERENCES

1. H.N Ting, Y. Jasmy and S.H Salleh. "*Malay Syllable Recognition Using Neural Network".* In Proceeding of the International Student Conference on Research and Development, SCOReD, Kuala Lumpur, 2006
2. Abas Lutfi. *"Linguistik Deskriptif Nahu". Dewan Bahasa dan Pustaka*, Kuala Lumpur: pp. 10 - 20 (1971)
3. J.L Rouas, J. Farinas, F. Pellegrino and R.A Obrecht, "*Rhythmic Unit Extraction and Modeling for Automatic Language Identification*". Speech Communication 47: 436-456. 2005
4. Md Sah Hj Salam and Mohamad Nasir Said Ibrahin. "*An Initial Experiment on Syllable Based Approach For Malay Digit Recognition".* In Proceeding of Advance Technology Congress. ATC2003, Putrajaya, Malaysia 2003.
5. B. Michele and V.N. Igor, "*Detection of Abrupt Changes: Theory and Application*", *Prentice Hall, INC*. USA 1993

6. S. Jarifi, D. Pastor and O. Rosec,. *"Brandt's GLR Method & Refined HMM Segmentation for TTS Synthesis Application"*. In Proceeding of European Signal Processing Conference, EUSIPCO'2005. Antalya,Turkey. 2005
7. T. Jehan T. *"Musical Signal Parameter Estimation"*. Master Thesis, University of Rennes, France. 1997
8. R.A. Obrecht, "*Automatic Segmentation of Continuous Speech Signal"*, IEEE Trans. Acoustic, Speech and Signal Processing, vol ASSP-36(1). pp 29-40, 1988
9. K. Kohler *"Segmental reduction in connected speech in German: Phonological facts and phonetics explaination"*. Speech Production and Speech Modelling, Kluwer, Dordrecht. pp.69-92. 1990
10. O. Engstrand, *"Sytematicity of phonetic variation in natural discourse"*. Speech Communication 11, pp. 337-346. 1992
11. *Language Production and Perception*, online : http://www.ling.upenn.edu/courses/Fall_1998/ling001/production_perception.html. pp. 1 – 11.
12. P. Cosi, J.P. Hosom, and F. Tesser. "*High performance Italian continuous digit recognition*", In Proceedings of International Conference on Spoken Language Processing, Beijing, China, ICSLP 2000.
13. W. Wei .and and S.V. Vuuren. *"Improved neural Network Training of Inter-Word Context Units for Connected Digit recognition"*, In Proceeding of IEEE International Conf. on Acoustics, Speech & Signal Processing, Seattle, ICASSP 1998
14. T. Nuttakorn and K. Boonserm. " *A syllable - based connected Thai digit speech recognition using neural network and duration modeling"*. In Proceeding of *The 1999 IEEE International Symposium on Intelligent Signal Processing and Communication. Pp 785-788. 1999.*
15. L.R Rabiner and M.R Sambur. "*Some Preliminary Experiments in the Recognition of Connected Digits".* IEEE Trans. Acoustic, Speech and Signal Processing, vol ASSP-24. pp 170-182 April 1976
16. Md Sah Hj Salam , Dzulkifli Mohamad dan S-H Salleh." *Speech Anticipation via Genetic Optimization: An Experiment on Simulated Data",* In Proceeding of International .Conference on Artificial Intelligence in Engineering and Technology, ICAIET '06, Kota Kinabalu, Sabah, Malaysia.2006.