

Evidences of correspondences

Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation

- 1. First Submission (01 October 2013)**
2. Second Submission, Respond to Reviewers (02 April 2014)
3. Final Submission, Respond to Reviewers (26 May 2014)



Article Summary

Title	Automatic Segmentation of Speech into Syllabic Units Using Fuzzy Smoothed Local Normalized Energy Contour				
Author(s)	Mr. Suyanto Suyanto, Agfianto Eko Putra				
Series	C : Information and Communication Technology				
Abstract	<p>This paper discusses the usage of short term energy contour of a speech smoothed by a fuzzy-based method to automatically segment the speech into syllabic units. Two additional procedures, local normalization and postprocessing, are proposed to improve the method. Testing to Indonesian speech dataset shows that local normalization significantly improves the accuracy of fuzzy smoothing. In postprocessing step, the procedure of splitting missed short syllables reduces the deletion errors, but unfortunately it increases the insertion ones. On the other hand, an assimilation of a single consonant segment into its previous or next segment reduces the insertion errors, but increases the deletion ones. The sequential combination of splitting and then assimilation gives quite significant improvement of accuracy as well as reduction of deletion errors, but it slightly increases the insertion ones.</p>				
Keywords	<i>assimilation of single consonant segment, fuzzy-based smoothing; local normalization; short term energy contour; speech segmentation; splitting missed short syllables, syllabic units</i>				
Status	Review In Progress				
History	<table><tr><td>1st Submission (01 October 2013)</td></tr><tr><td>Article File : C13363-01.doc</td></tr><tr><td>Reviewer #1 :<ul style="list-style-type: none">• Evaluation Result :• Comment :</td></tr><tr><td>Reviewer #2 :<ul style="list-style-type: none">• Evaluation Result :• Comment :</td></tr></table>	1st Submission (01 October 2013)	Article File : C13363-01.doc	Reviewer #1 : <ul style="list-style-type: none">• Evaluation Result :• Comment :	Reviewer #2 : <ul style="list-style-type: none">• Evaluation Result :• Comment :
1st Submission (01 October 2013)					
Article File : C13363-01.doc					
Reviewer #1 : <ul style="list-style-type: none">• Evaluation Result :• Comment :					
Reviewer #2 : <ul style="list-style-type: none">• Evaluation Result :• Comment :					

:: Author Menu ::

- [Submit New Article](#)
- [View Submitted Articles](#)
- [View Message Box](#)
- [Change Profile](#)
- [Change Password](#)
- [Logout](#)



Automatic Segmentation of Speech into Syllabic Units using Fuzzy Smoothed Local Normalized Energy Contour

Suyanto¹, Agfianto Eko Putra²

¹School of Engineering, Telkom University

Jalan Telekomunikasi Terusan Buah Batu, Bandung 40257, Indonesia

²Faculty of Mathematics and Natural Sciences, Gadjah Mada University

Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia

Email: suyanto.s3.ilkomp@mail.ugm.ac.id, agfi@ugm.ac.id

Abstract. This paper discusses the usage of short term energy contour of a speech smoothed by a fuzzy-based method to automatically segment a clean continuous speech into syllable units. Two additional procedures, local normalization and postprocessing, are proposed to improve the method. Testing to Indonesian speech dataset shows that local normalization significantly improves the accuracy of fuzzy smoothing. In postprocessing step, the procedure of splitting missed short syllables reduces the deletion errors, but unfortunately it increases the insertion ones. On the other hand, the postprocessing step of assimilation of single consonant segments to either its previous or next segment, reduces the insertion errors but increases the deletion ones. The combination of splitting and assimilation gives quite significant improvement of accuracy as well as reduction of deletion error, but it slightly increases the insertion errors.

Keywords: *assimilation of single consonant segments, fuzzy-based smoothing; local normalization; short term energy contour; speech segmentation; splitting missed short syllables, syllabic units.*

1 Introduction

Information about syllabic units could be used to improve the performance of automatic speech recognition (ASR). Janakiraman in [1] reported that incorporating information of syllable boundaries into an ASR reduced both computational complexity and word error rate (WER) significantly compared to the flat start ASR. The WER could be reduced from 13% into 4.4% and 36% into 21.2% for TIMIT and NTIMIT databases respectively [1].

Segmentation of speech into syllabic units could be approached using three different features, i.e. time domain, frequency domain, and combination of them. Sheikhi and Almasganj in [2] reported that time domain feature based on fuzzy smoothed short term energy (STE) contour could be used to segment

¹Suyanto is now a doctoral student in Faculty of Mathematics and Natural Sciences, Gadjah Mada University

speech into syllabic units with accuracy of 93.8% for Farsi speech dataset, where the structure of Farsi syllable is CV(C)(C) with C is consonant and V is vowel. Unfortunately, this method produced high insertion error since syllables ending with nonstop nasal consonants such as /n/ or /m/ usually have two energy peaks.

In this research, the method in [2] is adapted to an Indonesian speech dataset, i.e. clean speech corpus as in [3] that contains more complex structures of Indonesian syllables, such as CCVC and CVCCC, and as described in [4]. Some modifications as well as additional procedures are proposed to improve the performance.

2 The proposed syllabification

This proposed method for automatic segmentation of speech into syllabic units (ASSIS) exploits STE contour smoothed by a fuzzy-based method with two additional procedures, i.e. local normalization and postprocessing, as illustrated by figure 1.

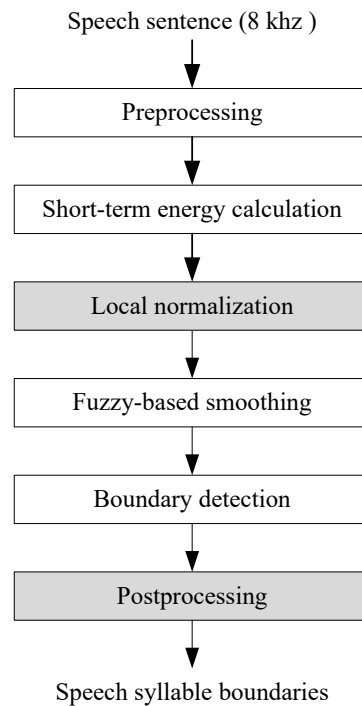


Figure 1 The block diagram of ASSIS.

2.1 Preprocessing

To spectrally flatten the signal, a pre-emphasis procedure is performed using equation 1, where α is the pre-emphasis coefficient, set to 0.9. As the speech signal sampled at a highly enough rate, samples of the low frequency tend to change slowly. Those such samples could be removed by subtracting a sample with the previous sample as described in this equation. In other words, the subtraction preserves samples that change rapidly, i.e. its high-frequency components.

$$y_i = x_i - \alpha x_{i-1} \quad (1)$$

Next, the emphasized signal is blocked into frames using Hamming windows, where each frame is 10 mili second containing 80 samples as frequency sampling used here is 8 khz. The frames are averlap of 60 samples (75% of the frame) to get smooth features.

A long sentence speech commonly contains some so long silences that it is quite hard to find the accurate syllable boundaries in the such speech. Hence, a threshold-based procedure of silence removal is performed by using both energy and duration thresholds on the STE. So, only the speech, with no silence, will be processed in the next step. At the final step, the removed silences will be restored to get the original speech.

2.2 Short-term energy

Short-term energy (STE) could be produced using some different formulas, such as absolute, square, root mean square, Teager and modified Teager. Sheikhi and Almasganj in [2] showed that the Teager energy gave the best accuracy for Farsi speech dataset. But, in this research, the square energy as in equation 1 is used as it empirically gives the best accuracy for Indonesian speech dataset. This formula increases the difference between low signal energy and the higher one.

$$E = \sum_{i=1}^N S_i^2 \quad (1)$$

2.3 Local normalization

A long sentence speech could contain high amplitudes in some parts and low amplitudes in other parts. A local normalization is performed to the STE contour by detecting frames containing very low energy and then the set of high energy frames occured between the very low energy are normalized to the

maximum energy in the set. This step produces a better STE contour for fuzzy based-smoothing to easily tract the contour.

2.4 Fuzzy-based smoothing

The local normalized STE is then smoothed based on the 7 previous energy samples (E_1, E_2, \dots, E_7) using a fuzzy-based smoothing as in described [2]. But, the fuzzy linguistic rules are modified to have 11 rules (instead of 7 rules) to cover varying crisp value inputs $x_i = E_i - \hat{E}_i$, i.e. energy of speech subtracted by the fuzzy smoothed energy, as listed in table 1.

Table 1 The fuzzy linguistic rules.

No	Fuzzy linguistic rules
1	if most inputs are very small positif then output is very small positif
2	if most inputs are small positif then output is small positif
3	if most inputs are medium positif then output is medium positif
4	if most inputs are big positif then output is big positif
5	if most inputs are very big positif then output is very big positif
6	if most inputs are very small negatif then output is very small negatif
7	if most inputs are small negatif then output is small negatif
8	if most inputs are medium negatif then output is medium negatif
9	if most inputs are big negatif then output is big negatif
10	if most inputs are very big negatif then output is very big negatif
11	else output is zero

Membership functions for any rule and term *most* in fuzzy linguistic rules as well as the activity degree of any fuzzy group are adapted from [2]. The membership functions for all fuzzy rules are described by equation 2, where A is a fuzzy rule, c_A is the center point of A 's membership function and w is the width of membership function. In this research, all membership functions have the same width and $w/2$ overlap, where $w = 0.18$ is found from some experiments. The center point of 11-th fuzzy rule is zero.

$$\mu_A(x_i) = \begin{cases} \frac{-2(x_i - c_A)}{w+1} & c_A - \frac{w}{2} < x_i < c_A \\ \frac{2(x_i - c_A)}{w+1} & c_A < x_i < c_A + \frac{w}{2} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The term *most* in fuzzy linguistic rules is defined by equation 3.

$$\mu_{most}(z) = \begin{cases} 0 & z \leq 0.1 \\ 0.5 \left(1 - \cos \left[\frac{\pi(z-0.1)}{0.8} \right] \right) & 0.1 < z < 0.9 \\ 1 & z \geq 0.9 \end{cases} \quad (3)$$

The activity degree of any fuzzy group is described by equation 4.

$$\lambda_A = \text{median}[\mu_A(x_i) : x_i \in A] * \mu_{most} \left[\frac{\text{number of } x_i \in A}{\text{total number of } x_i} \right] \quad (4)$$

Output of the fuzzy-based smoothing is a correlation product in equation 5.

$$\Delta E = \sum_{A=1}^{11} c_A \lambda_A \quad (5)$$

Finally, the fuzzy smoothed energy is calculated by equation 6.

$$\hat{E}_{i+1} = \hat{E}_i + \Delta E \quad (6)$$

2.5 Boundary detection

The threshold method based on local minima detection as proposed by Sheikhi and Almasganj in [2] is adapted in this research. There are three parameters should be tuned carefully, i.e. D_1 , frame duration on the right and left of an energy sample to decide the sample as a maximum energy point; Th , the threshold for ratio of a maximum energy point and a consecutive minimum energy point to decide the point as a local maximum; and D_2 , frame duration to decide a local minimum energy point as syllable boundary or not. From some observations on Indonesian speech dataset, optimum values for those parameters are found, i.e. $D_1 = 3$, $Th = 1.5$, and $D_2 = 20$;

Figure 2 illustrates the segmentation of an Indonesian utterance “*Dengan skema ini*” (By this scheme) using fuzzy-based smoothing for both global and local normalized STE. In global normalized STE, two syllables /i/ and /ni/ in the utterance produce so low energies that they are flat after fuzzy smoothing and recognized as one syllable /ini/. On the other hand, the fuzzy smoothed local normalized STE gives a better contour for the boundary detection procedure to accurately produce 6 syllables, /de/-/ngan/-/ske/-/ma/-/i/-ni/, as performed by a linguistic expert.

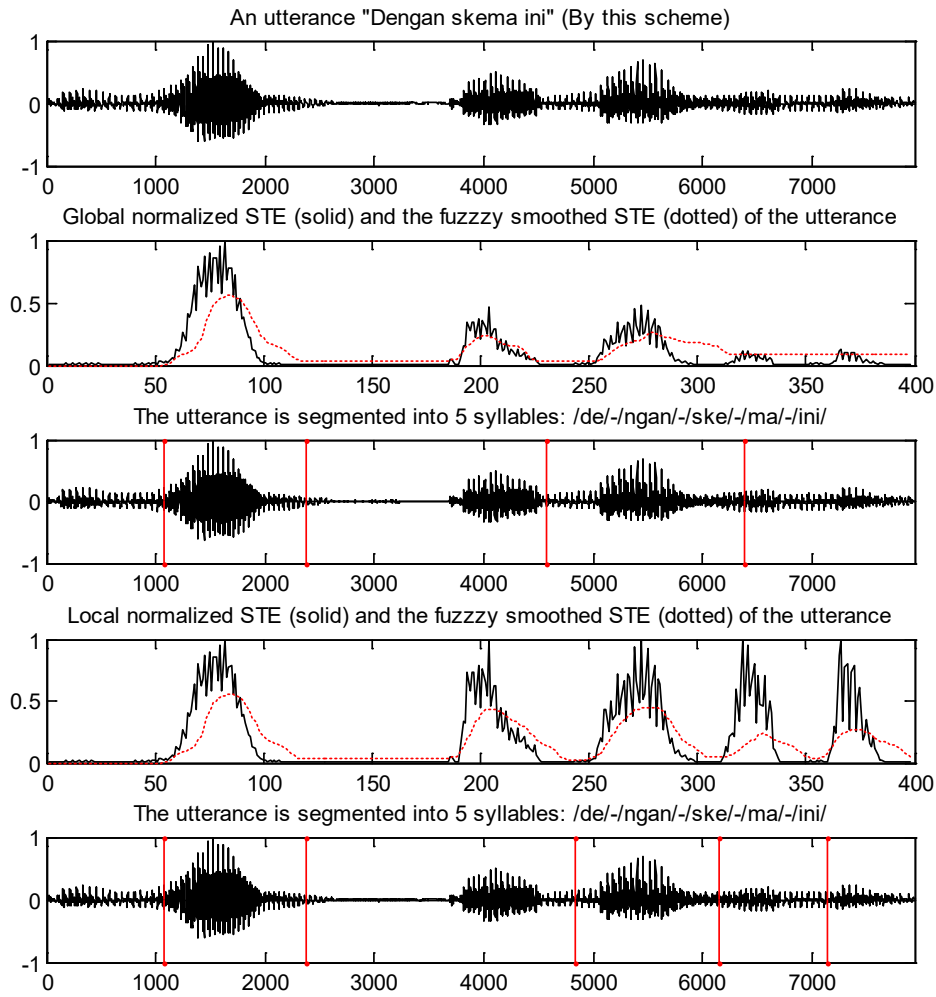


Figure 2 An Indonesian utterance “*Dengan skema ini*” (1), global normalized STE and the fuzzy smoothed one (2), and the segmentation result (3), local normalized STE and the fuzzy smoothed one (4) and the segmentation result (5).

2.6 Postprocessing

Indonesian two consecutive syllables producing vowel series, i.e. the first syllable ending with a vowel and the second one has a vowel on the beginning, commonly have a single energy peak so that they produce deletion error in syllable detection. Hence, a threshold-based splitting procedure is performed to split those syllables. First, the syllable segments produced by the previous step are scanned and the STE of each segment is recalculated using lower frame size of 9 ms (instead of 10 ms) to find more significant variation of energy. An

energy valley could be a syllable boundary if energy ratio and duration between its lowest neighbor peak and the valey as well as both parameters regarding its highest neighbor peak are bigger than defined thresholds.

As Farsi syllables stated in [2], Indonesian syllables ending with nonstop nasal consonants such as /m/ and /n/ as well as high energy unvoiced consonants /h/ and glottal stop, usually have two energy peaks that cause insertion error. A further observation to fricative consonants /f/, /s/, /z/, /sy/, and /kh/ as in [5] shows that they also cause the such error. Hence, an assimilation procedure as described in [6] is adapted to delete the unexpected boundaries. But, in this research, both total and residual energies are calculated using square energy (instead of log root square energy as in [6]) of original and the low pass filtered signal respectively, with frame size of 10 ms. In figure 3, the consonant segments /s/, /f/, and /c/ produce residual energies those are significantly lower than the total ones, but the other segments yield residual those are very similar to their total energies. This fact is exploited to **assimilate** the such consonant segments to their neighbors. The three thresholds to decide the assimilation as described in [6], i.e. *MaxRatio*, *AverageRatio* and *DecreasingResidualRatio*, are defined empirically by some observations.

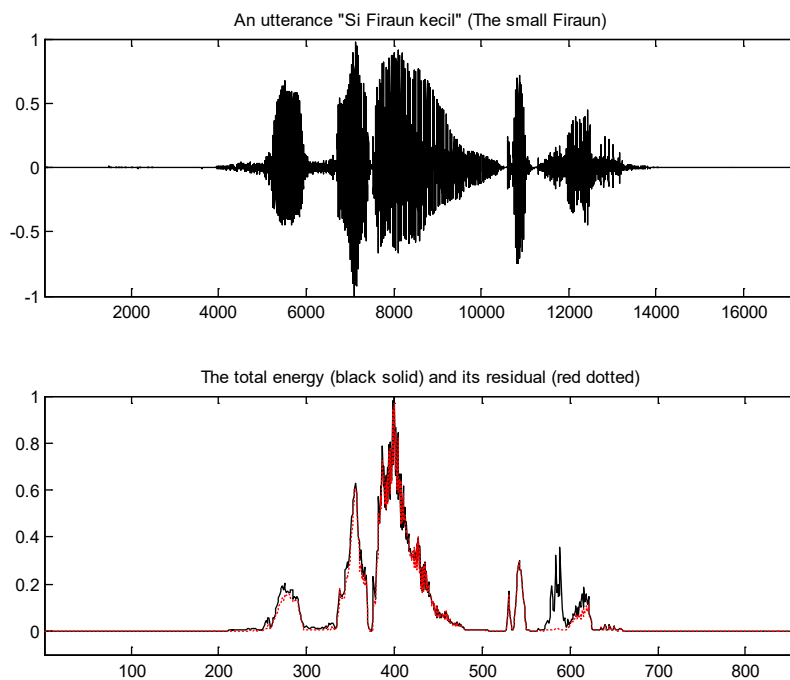


Figure 3 An utterance “*Si Firaun kecil*” (The small Firaun), the total energy (black solid line) and its residual energy (red dotted).

3 Result and Discussion

Speech dataset used here is taken from the Indonesian phonetically balanced speech corpus as described in [3]. The corpus contains 44,000 utterances from 400 speakers, where each speaker reads 110 sentences. This research takes 220 utterances from two speakers, a male and a female. Those utterances cover varying structures of syllable, from the simple ones such as V, CV, and CVC, to the complex ones such as CCVC and CVCCC, as described in [4].

To see the performance of proposed local normalization and post processing, five different ASSIS systems are developed, i.e. ASSIS with global normalization (AGN), ASSIS with local normalization (ALN), ALN with splitting (ALNS), ALN with assimilation (ALNA), and ALN with splitting and assimilation (ALNSA). Testing to the speech corpus of 220 speeches containing 3,360 syllables gives results listed in table 2, where accuracy is defined as percentage of detected syllables with less than 50 mili second error (i.e. 30% of average syllable duration in the speech corpus).

Table 2 The performance of AGN, ALN, ALNS, ALNA, and ALNSA.

Type of ASSIS	Accuracy	Insertion (%)	Deletion (%)
AGN	66.26	8.81	20.61
ALN	82.53	8.49	5.85
ALNS	86.37	9.99	3.34
ALNA	81.79	7.51	7.04
ALNSA	85.57	8.93	4.11

Compare to AGN, the proposed ALN gives significantly better performance: improves the accuracy up to 16.27%, reduces deletion errors of 14.76%, and slightly reduces the insertion errors of 0.32%. The local normalization works very well and the thresholds could be found easily since the corpus used here is clean read speech.

ALNS improves accuracy up to 4.16% and reduces deletion errors of 2,51%, but it increases insertion errors of 1.50 %. This is a proof that splitting procedure is capable to detect short segments. But, in some cases, it oversplits causing the insertion errors increase. It is quite hard to find the optimum threshold for this procedure. ALNA reduces insertion error of 0.98%, but it increases the deletion error of 1.19%. This shows that assimilation procedure, as it was designed, works only on any single consonant segment. It does not work on very long segments that need a splitting, instead of an assimilation. ALNSA reduces insertion error of 0.98%, but it increases the deletion error of 1.19%. The assimilation procedure.

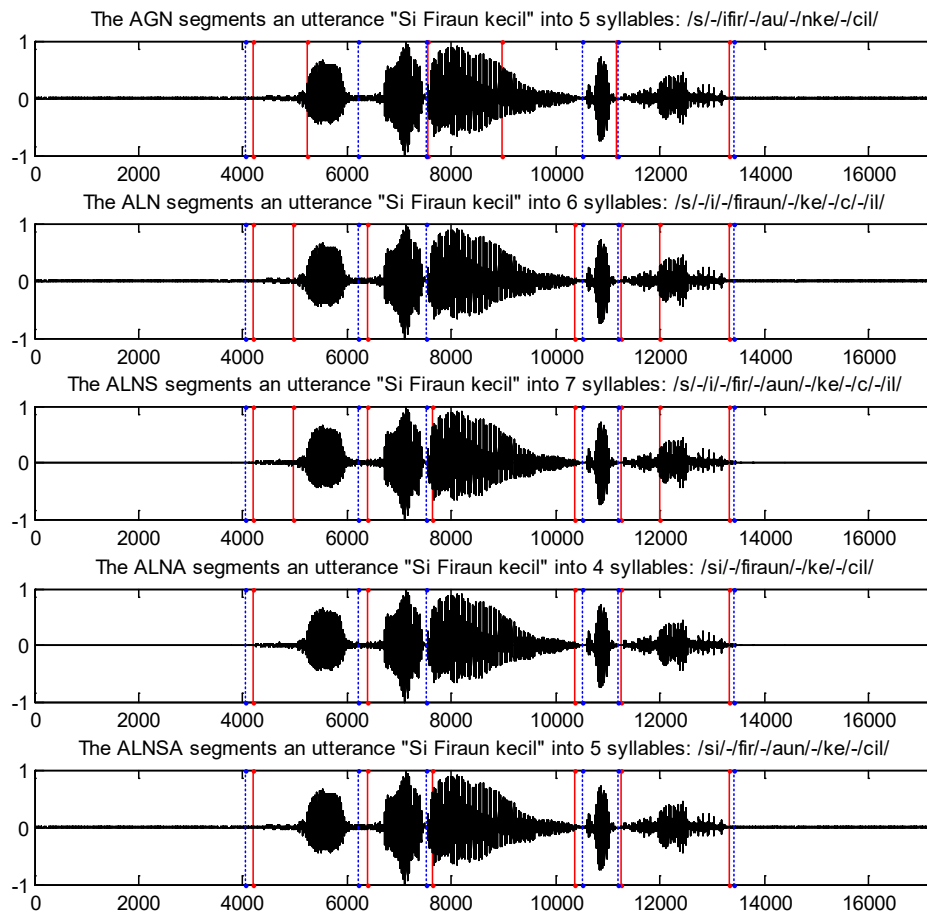


Figure 1 Segmentation of an utterance "Si Firaun kecil" (The small Firaun) using AGN, ALN, ALNS, ALNA, and ALNSA.

Figure 3 describes how those ASSIS systems work. An utterance "Si Firaun kecil" (The small Firaun) should be segmented into 5 syllables, /si/-/fir/-/aun/-/ke/-/cil/, where the boundaries are shown by dotted blue lines. AGN segments the utterance into 5 syllables, /s/-/ifir/-/au/-/nke/-/cil/, shown by solid red lines. It produces two insertion errors, i.e. boundaries between /s/ and /i/ and between /au/ and /n/, and also two deletion errors, i.e. syllable boundaries between /si/ and /fir/ and between /aun/ and /ke/. ALN segments the utterance into 6 syllables, /s/-/i/-/firaun/-/ke/-/c/-/il/. It produces two insertions, where single consonant segments /s/ and /c/ should be assimilated into their right segments, and a deletion, i.e. /firaun/ should be split into /fir/ and /aun/. ALN removes one insertion and two deletions produced by AGN. But, it produces a new insertion, between /c/ and /il/, and a new deletion between /fir/ and /aun/. ALNS segments

the utterance into 7 syllables, /s/-/i/-/fir/-/aun/-/ke/-/c/-il/. It produces similar segments as those by ALN, but the deletion between /fir/ and /aun/ could be removed. It shows splitting procedure works accurately. ALNA segments the utterance into 4 syllables, /si/-/firaun/-/ke/-/cil/. It removes two insertions produced by ALN. It shows the assimilation procedure works accurately to merge a single consonant segment into the expected segment. ALNA can not remove the deletion between /fir/ and /aun/ since the assimilation procedure is not designed to split a long segment. ALNSA accurately segments the utterance into 5 syllables, /si/-/fir/-/aun/-/ke/-/cil/. It removes two insertions as well as a deletion produced by ALN. It shows both splitting and assimilation procedures work very well.

4 Conclusion

The proposed local normalization significantly improves the performance of fuzzy-based smoothing, i.e. increasing accuracy as well as reducing insertion and deletion errors. Two postprocessing procedures, splitting and assimilation, work in different ways. The procedure of splitting missed short syllables reduces the deletion errors, but unfortunately it increases the insertion ones. On the other hand, the assimilation of single consonant segments into either its previous or next segment, reduces the insertion errors but increases the deletion ones. The combination of splitting and assimilation gives quite significant improvement of accuracy as well as reduction of deletion error, but it slightly increases the insertion errors.

Acknowledgement

Many thanks to all colleagues in Faculty of Mathematics and Natural Sciences, University of Gadjah Mada and also all colleagues in School of Engineering, Telkom University (former: Telkom Institute of Technology) for the supports.

References

- [1] Janakiraman, R., Kumar, J.C., and Murthy, H.A., *Robust syllable segmentation and its application to syllable-centric continuous speech recognition*, in Proceedings of National Conference on Communications (NCC), pp. 1-5, 2010.
- [2] Sheikhi, G. & Almasganj, F., *Segmentation of speech into syllable units using fuzzy smoothed short term energy contour*, in Proc. the 18th Iranian Conference on BioMedical Engineering, 14-16 December 2011, Tehran, Iran, pp. 195-198, 2011.

- [3] Suyanto & Adityatama, J., 2012, *Yooi: An Indonesian Short Message Dictation*, International Journal of Intelligent Information Processing (IJIP), Republic of Korea, **3**(4), pp. 68-74, 2012.
- [4] Suyanto & Hartati, S., *Design of Indonesian LVCSR Using Combined Phoneme and Syllable Models*, in Proceedings of the 7th International Conference on Information & Communication Technology and Systems (ICTS), Bali, Indonesia, pp. 191-196, 2013.
- [5] Alwi, H., Dardjowidjojo, S., Lapoliwa, H., and Moeliono, A.M., *Tata bahasa baku bahasa Indonesia (The standart Indonesian grammar)*, Jakarta, Balai Pustaka, 1998.
- [6] Petrillo, M. & Cutugno, F., *A syllable segmentation algorithm for English and italian*, in Proceedings of EUROSPEECH, pp. 2913-2916, 2003.

Evidences of correspondences

Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation

1. First Submission (01 October 2013)
2. Second Submission, Respond to Reviewers (02 April 2014)
3. Final Submission, Respond to Reviewers (26 May 2014)



Article Summary

Title	Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation
Author(s)	Mr. Suyanto Suyanto, Agfianto Eko Putra
Series	C : Information and Communication Technology
Abstract	This paper discusses the usage of short term energy contour of speech smoothed by a fuzzy-based method to automatically segment a speech into syllabic units. Two new additional procedures, local normalization and post-processing, are proposed to adapt to the Indonesian language. Testing to 220 Indonesian utterances shows that the local normalization significantly improves the performance of fuzzy-based smoothing. In the post-processing procedure, splitting and assimilation work in different ways. Splitting missed short syllables sharply reduces the deletion, but slightly increases the insertion. On the other hand, assimilation of a single consonant segment into an expected previous or next segment slightly reduces the insertion, but increases the deletion. The use of splitting procedure gives higher accuracy than the assimilation procedure and the combined splitting-assimilation since, in many cases, the assimilation keeps unexpected insertions and over merges expected segments.
Keywords	<i>assimilation, fuzzy-based smoothing; Indonesian language; local normalization; short term energy contour; splitting; syllable segmentation</i>
Status	Revised
History	2nd Submission (02 April 2014)

Article File : C13363-02.doc

Author's Comment : commC13363-02.docx

Reviewer #1 :

- **Evaluation Result :**
- **Comment :**







Reviewer #2 :

- **Evaluation Result :**
- **Comment :**

Reviewer #3 :

- **Evaluation Result :**
- **Comment :**

:: Author Menu ::

-  [Submit New Article](#)
-  [View Submitted Articles](#)
-  [View Message Box](#)
-  [Change Profile](#)
-  [Change Password](#)
-  [Logout](#)

1. Reviewers' comments:

The number of references is ok now, but the related work section is still not elaborate enough (the difference between related work and introduction sections is not obvious). I highly suggest that authors move some of the introduction paragraphs into related works section (those that discuss state-of-the-art of methods). Additionally, more paragraphs need to be added in the Introduction section, providing the context (background) from which the main problem could occur, why it is important and the motivation behind solving the problem. The last paragraph in the introduction should describe the paper organization.

Revision by authors:

>> We have moved some of the introduction paragraphs into related works section as follows:

2. Related works

Segmentation of speech into syllabic units can be approached using three different features, i.e. 1) time domain, such as in [12], [15], [16], [17]; 2) frequency domain, such as in [11], [18], [19], [20], [21], [22], [23], [24], [25]; and combination of them, such as in [26], [27], [28]. The time domain approaches commonly use short-term energy (STE) smoothed by a smoothing algorithm, but the frequency domain approaches exploit cepstrum features.

A time domain approach in [16] simply uses a plain STE contour and a threshold to detect the locations of starting and end of syllable. This method works very well only for short-sentence utterances. In [12], the plain STE contour is firstly smoothed by fuzzy-based smoothing before defining the syllable boundaries using a threshold-based method. The usage of fuzzy-based smoothing gives much higher accuracy, 93.8% for Farsi speech dataset, than common moving average smoothing method. Unfortunately, this method produced high insertion error, i.e. 14.2%, since syllables ending with nonstop nasal consonants such as /n/ or /m/ usually have two energy peaks.

The frequency domain approaches are dominated by exploiting minimum phase group delay function [18], [19]. Those methods are improved significantly by incorporating a procedure called vowel onset point (VOP) detection which is capable of decreasing deletion and insertion errors as discussed in [11].

Compare to the frequency domain approach, the time domain approach generally faster, but unfortunately it produces more deletion and insertion errors. However, those errors can be reduced by performing frequency-based postprocessing procedure as described in [26] or by incorporating the VOP detection.

>> We have revised the Introduction section by adding some paragraphs providing the context (background) from which the main problem could occur, why it is important and the motivation behind solving the problem, and the paper organization as follows:

1. Introduction

Information about syllabic units can be used to improve the performance of flat start-based automatic speech recognition (ASR) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. In 2010, Janakiraman et. al. [11] reported that incorporating information of syllable boundaries into an English ASR reduced both computational complexity and word error rate (WER) significantly compared to the flat start ASR. The WER can be reduced from 13% to 4.4% and from 36% to 21.2% for TIMIT and NTIMIT databases respectively.

Each language has unique characteristics. For example, English and Indonesian language have different syllable patterns. A study to Telephone Conversations and Switchboard Corpus by Su-Lin Wu [3] shows that English has 80% monosyllabic words and 85% of them are simple structures (V, VC, CV, CVC) and the rest are complex structures such as CCCVC or CVCCC, where C is consonant and V is vowel. Our exploration to around 50 thousand words from the great dictionary of Indonesian language (*Kamus Besar Bahasa Indonesia* or KBBI) fourth edition, released in 2008 by *Pusat Bahasa*, shows that Indonesian language has only 1.57% monosyllabic words but it has much more simple structured syllables, up to 98.60%, than English. Hence, an Indonesian ASR is better developed using syllabic units with [syllable segmentation as an important sub system for the ASR](#).

This research focuses on syllable segmentation for Indonesian language. A segmentation method in [12] which is designed for Farsi language with simple syllable structures, CV(C)(C), is adapted and tested to Indonesian speech dataset of clean speech corpus as in [13]. Some modifications as well as two additional procedures, i.e. local normalization and postprocessing, are proposed to adapt to the Indonesian language which has some complex syllable structures, such as CCVC and CVCCC as described in [14].

The rest of this paper is organized as follows: [section 2 discusses the related works to syllable segmentation for some languages](#), [section 3 describes the proposed Indonesian syllable segmentation](#), [section 4 reports experimental result and discussion](#), and finally [section 5 gives some conclusions](#).

2. Reviewers' comments:

The second revision is improving the figure readability. In the current version, the use of thin, gray line in Figures 3,4 & 5 is susceptible to make the graphics lines unreadable when printed on black & white color. For revision, use black line, enlarge the figure's graphics and remove part the graphics that are empty (are not necessary).

Revision by authors:

[We have upgraded the figure using our simulaion to generate new figures with black and white instead of gray line](#), [enlarged the figures](#), and [removed part the graphics that are not necessary](#), as follows:

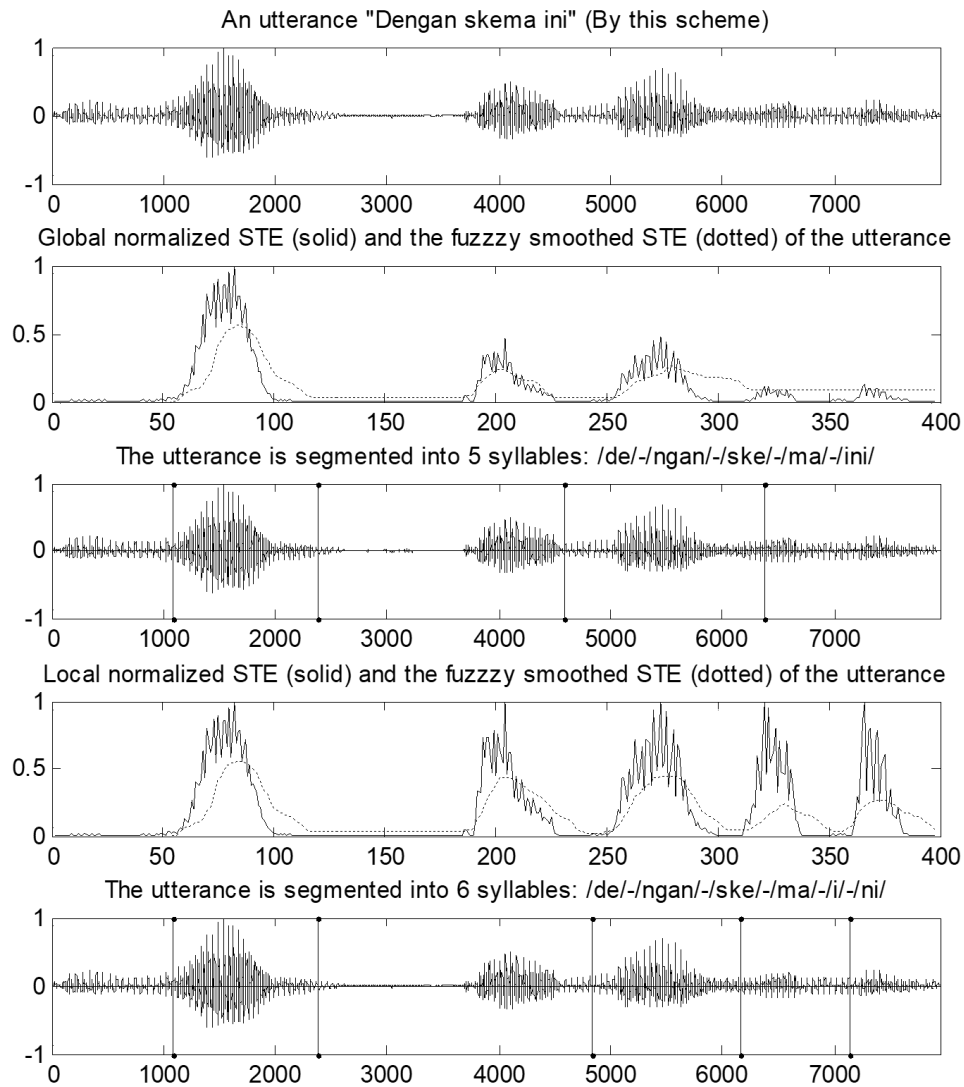


Figure 2 An Indonesian utterance “*Dengan skema ini*” (1), global normalized STE—solid line—and the fuzzy smoothed STE—dotted line (2), the segmentation boundaries produced using global normalized STE (3), local normalized STE and the fuzzy smoothed one (4), the segmentation boundaries produced using local normalized STE (5).

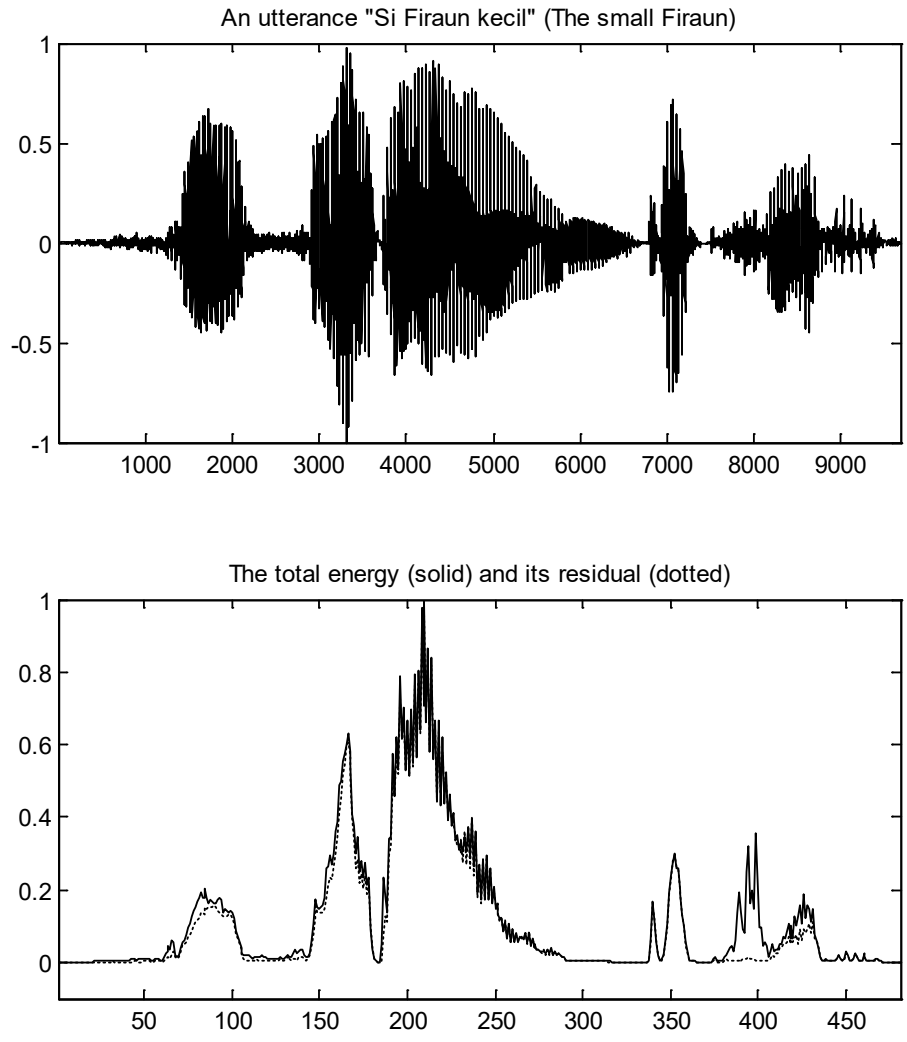


Figure 3 An utterance “Si Firaun kecil” (The small Firaun) and the total energy (solid line) as well as its residual energy (dotted line).

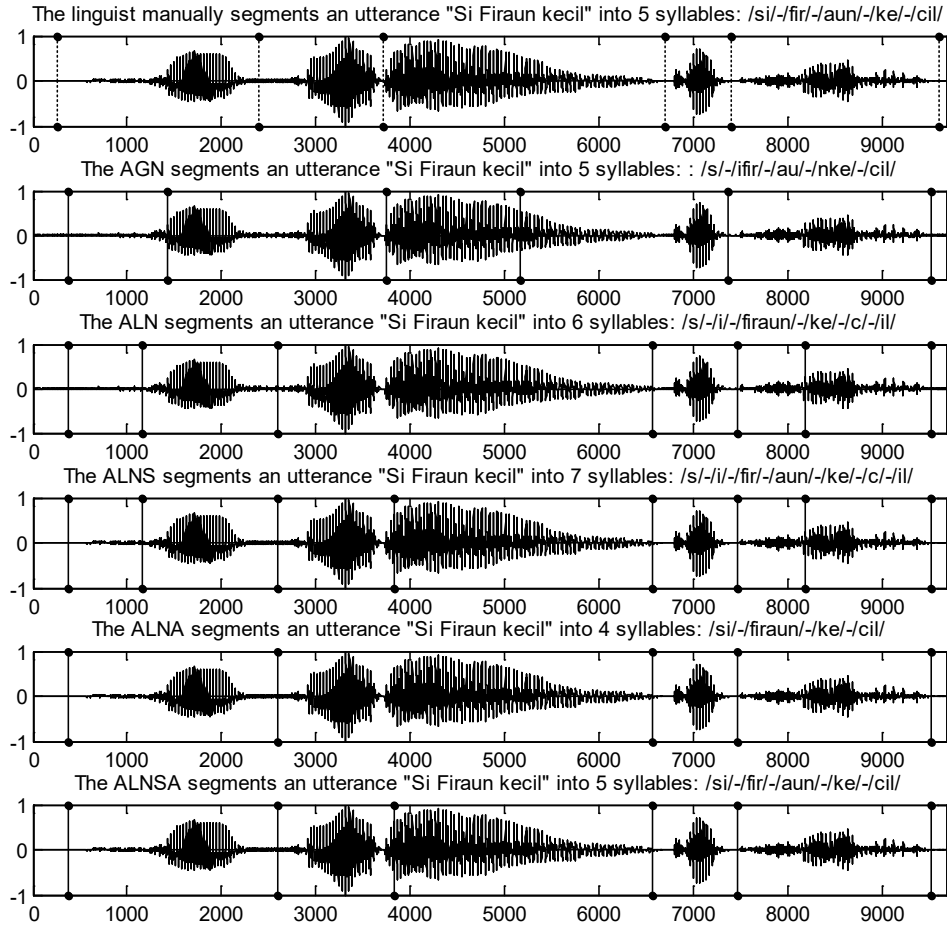


Figure 4 Segmentation of an Indonesian utterance, “Si Firaun kecil” (The small Firaun), by the linguist, AGN, ALN, ALNS, ALNA, and ALNSA.

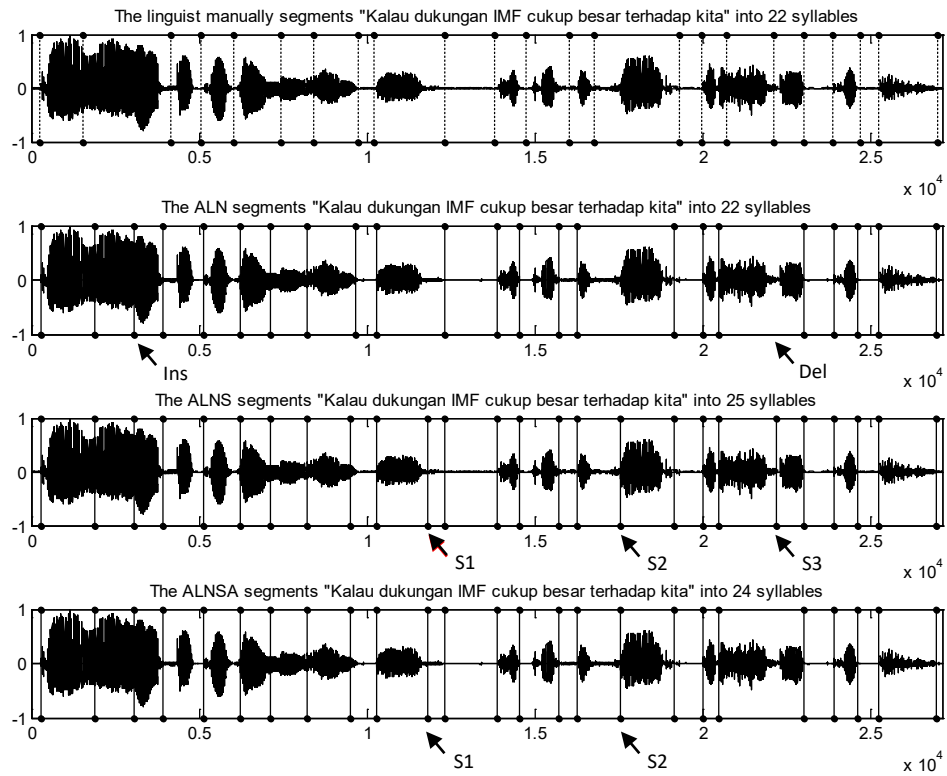


Figure 5 Segmentation of an Indonesian utterance, “Kalau dukungan IMF cukup besar terhadap kita” (If support of IMF is quite big to us), by the linguist, ALN, ALNS, and ALNSA.

3. Reviewers' comments:

I still found grammatical error here and there. Proof-read the paper again before submit the final revision to make sure that it has free grammatical error.

Revision by authors:

Yes, we have proof-read the paper and made some corrections. We believe that there is no grammatical error.

Evidences of correspondences

Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation

1. First Submission (01 October 2013)
2. Second Submission, Respond to Reviewers (02 April 2014)
3. Final Submission, Respond to Reviewers (26 May 2014)

Automatic Segmentation of Indonesian Speech into Syllables using Fuzzy Smoothed Energy Contour with Local Normalization, Splitting, and Assimilation

Suyanto¹, Agfianto Eko Putra²

¹School of Computing, Telkom University

Jalan Telekomunikasi Terusan Buah Batu, Bandung 40257, Indonesia

²Faculty of Mathematics and Natural Sciences, Gadjah Mada University

Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia

Email: suy@ittelkom.ac.id, agfi@ugm.ac.id

Abstract. This paper discusses the usage of short-term energy contour of speech smoothed by a fuzzy-based method to automatically segment a speech into syllabic units. Two new additional procedures, local normalization and postprocessing, are proposed to adapt to the Indonesian language. Testing to 220 Indonesian utterances shows that the local normalization significantly improves the performance of fuzzy-based smoothing. In the postprocessing procedure, splitting and assimilation work in different ways. Splitting missed short syllables sharply reduces the deletion, but slightly increases the insertion. On the other hand, assimilation of a single consonant segment into an expected previous or next segment slightly reduces the insertion, but increases the deletion. The use of splitting procedure gives higher accuracy than the assimilation and the combined splitting-assimilation procedures since, in many cases, the assimilation keeps the unexpected insertions and overmerges the expected segments.

Keywords: *assimilation, fuzzy-based smoothing; Indonesian language; local normalization; short-term energy contour; splitting; syllable segmentation.*

1 Introduction

Information about syllabic units can be used to improve the performance of flat start-based automatic speech recognition (ASR) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. In 2010, Janakiraman et. al. [11] reported that incorporating information of syllable boundaries into an English ASR reduced both computational complexity and word error rate (WER) significantly compared to the flat start ASR. The WER can be reduced from 13% to 4.4% and from 36% to 21.2% for TIMIT and NTIMIT databases respectively.

Each language has unique characteristics. For example, English and Indonesian language have different syllable patterns. A study to Telephone Conversations and Switchboard Corpus by Su-Lin Wu [3] shows that English has 80% monosyllabic words and 85% of them are simple structures (V, VC, CV, CVC)

and the rest are complex structures such as CCCVC or CVCCC, where C is consonant and V is vowel. Our exploration to around 50 thousand words from the great dictionary of Indonesian language (*Kamus Besar Bahasa Indonesia* or KBBI) fourth edition, released in 2008 by *Pusat Bahasa*, shows that Indonesian language has only 1.57% monosyllabic words but it has much more simple structured syllables, up to 98.60%, than English. Hence, an Indonesian ASR is better developed using syllabic units with syllable segmentation as an important sub system for the ASR.

This research focuses on syllable segmentation for Indonesian language. A segmentation method in [12] which is designed for Farsi language with simple syllable structures, CV(C)(C), is adapted and tested to Indonesian speech dataset of clean speech corpus as in [13]. Some modifications as well as two additional procedures, i.e. local normalization and postprocessing, are proposed to adapt to the Indonesian language which has some complex syllable structures, such as CCVC and CVCCC as described in [14].

The rest of this paper is organized as follows: section 2 discusses the related works to syllable segmentation for some languages, section 3 describes the proposed Indonesian syllable segmentation, section 4 reports experimental result and discussion, and finally section 5 gives some conclusions.

2 Related works

Segmentation of speech into syllabic units can be approached using three different features, i.e. 1) time domain, such as in [12], [15], [16], [17]; 2) frequency domain, such as in [11], [18], [19], [20], [21], [22], [23], [24], [25]; and combination of them, such as in [26], [27], [28]. The time domain approaches commonly use short-term energy (STE) smoothed by a smoothing algorithm, but the frequency domain approaches exploit cepstrum features.

A time domain approach in [16] simply uses a plain STE contour and a threshold to detect the locations of starting and end of syllable. This method works very well only for short-sentence utterances. In [12], the plain STE contour is firstly smoothed by fuzzy-based smoothing before defining the syllable boundaries using a threshold-based method. The usage of fuzzy-based smoothing gives much higher accuracy, 93.8% for Farsi speech dataset, than common moving average smoothing method. Unfortunately, this method produced high insertion error, i.e. 14.2%, since syllables ending with nonstop nasal consonants such as /n/ or /m/ usually have two energy peaks.

The frequency domain approaches are dominated by exploiting minimum phase group delay function [18], [19]. Those methods are improved significantly by

incorporating a procedure called vowel onset point (VOP) detection which is capable of decreasing deletion and insertion errors as discussed in [11].

Compare to the frequency domain approach, the time domain approach generally faster, but unfortunately it produces more deletion and insertion errors. However, those errors can be reduced by performing frequency-based postprocessing procedure as described in [26] or by incorporating the VOP detection.

3 The proposed syllable segmentation

The proposed automatic segmentation of Indonesian speech into syllables (ASISS) exploits STE contour smoothed by a fuzzy-based method with two additional procedures, i.e. local normalization and postprocessing, as illustrated by figure 1.

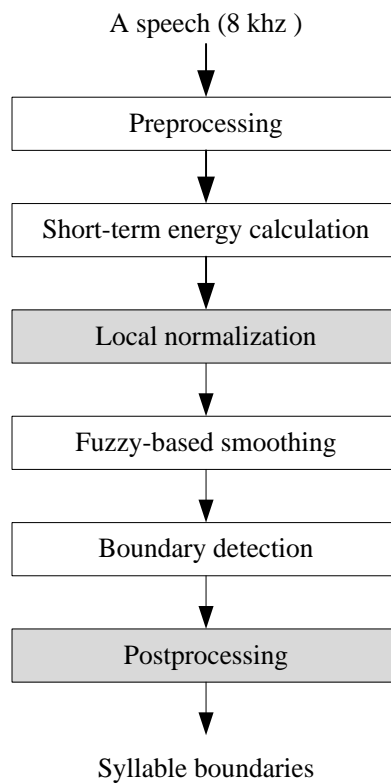


Figure 1 The block diagram of ASISS.

3.1 Preprocessing

To spectrally flatten the speech signal, a preemphasis procedure is performed using equation 1, where α is the preemphasis coefficient, set to 0.9. As the speech signal is sampled at a highly enough rate, the samples of the low frequency tend to change slowly. Those such samples can be removed by subtracting a sample with the previous sample as described in this equation. In other words, the subtraction preserves samples that change rapidly, i.e. its high frequency components.

$$y_i = x_i - \alpha x_{i-1} \quad (1)$$

Next, the emphasized signal is blocked into frames using Hamming windows, where each frame is 10 mili second (ms) containing 80 samples as frequency sampling used here is 8 khz. The frames are averlap of 60 samples, i.e. 75% of the frame, to get smooth features.

A long sentence speech commonly contains some so long silences that it is quite hard to find the accurate syllable boundaries in the such speech. Hence, a threshold-based procedure of silence removal is performed using both energy and duration thresholds on the STE. So, only the speech (no silence) will be processed in the next step. In the final step, the removed silences will be restored to reproduce the original speech.

3.2 Short-term energy

An STE can be produced using some different formulas, such as absolute, square, root mean square, Teager, and modified Teager. Sheikhi and Almasganj [12] shows that the Teager energy gives the best accuracy for Farsi speech dataset. But, in this research, a square energy as in equation 2 is used since it can increase the difference between low signal energy and the higher one and it empirically gives the best accuracy for Indonesian speech dataset.

$$E = \sum_{i=1}^N S_i^2 \quad (2)$$

3.3 Local normalization

A long sentence speech may contain high amplitudes in some parts and low in the others. Hence, a local normalization is performed to the STE contour by detecting frames of very low energy and then normalizing a set of high energy frames occurs between those very low energy frames into the maximum energy in that set. This step is expected to produce a better STE contour than that produced by global normalization used in [12].

3.4 Fuzzy-based smoothing

The local normalized STE is then smoothed based on the seven previous energy samples (E_1, E_2, \dots, E_7) using a fuzzy-based smoothing as described in [12]. But, the fuzzy linguistic rules are modified to have 11 rules (instead of 7 rules), as listed in table 1, to adapt the varying crisp valued inputs.

Membership functions for any rule and term *most* in fuzzy linguistic rules as well as the activity degree of any fuzzy group are adapted from [12]. The membership functions for all fuzzy rules are described by equation 3, where $x_i = E_i - \hat{E}_i$, i.e. crisp valued inputs come from speech energy subtracted by the fuzzy smoothed one, A is a fuzzy rule, c_A is the center point of A 's membership function, and w is the width of membership function [12]. In this research, all membership functions have the same width and $w/2$ overlap, where $w = 0.18$ is found by some experiments. The center point of the 11-th fuzzy rule is zero.

Table 1 The fuzzy linguistic rules.

No	Fuzzy linguistic rules
1	if most inputs are very small positif then output is very small positif
2	if most inputs are small positif then output is small positif
3	if most inputs are medium positif then output is medium positif
4	if most inputs are big positif then output is big positif
5	if most inputs are very big positif then output is very big positif
6	if most inputs are very small negatif then output is very small negatif
7	if most inputs are small negatif then output is small negatif
8	if most inputs are medium negatif then output is medium negatif
9	if most inputs are big negatif then output is big negatif
10	if most inputs are very big negatif then output is very big negatif
11	else output is zero

$$\mu_A(x_i) = \begin{cases} \frac{-2(x_i - c_A)}{w+1} & c_A - \frac{w}{2} < x_i < c_A \\ \frac{2(x_i - c_A)}{w+1} & c_A < x_i < c_A + \frac{w}{2} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The term *most* in fuzzy linguistic rules is defined by equation 4 [12].

$$\mu_{most}(z) = \begin{cases} 0 & z \leq 0.1 \\ 0.5 \left(1 - \cos \left[\frac{\pi(z-0.1)}{0.8} \right] \right) & 0.1 < z < 0.9 \\ 1 & z \geq 0.9 \end{cases} \quad (4)$$

The activity degree of any fuzzy group is described by equation 5 [12].

$$\lambda_A = \text{median}[\mu_A(x_i): x_i \in A] * \mu_{most} \left[\frac{\text{number of } x_i \in A}{\text{total number of } x_i} \right] \quad (5)$$

Output of the fuzzy-based smoothing is a correlation product in equation 6.

$$\Delta E = \sum_{A=1}^{11} c_A \lambda_A \quad (6)$$

Finally, the fuzzy smoothed energy is calculated by equation 7 [12].

$$\hat{E}_{i+1} = \hat{E}_i + \Delta E \quad (7)$$

3.5 Boundary detection

A threshold method based on local minima detection as proposed by Sheikhi and Almasganj in [12] is adapted in this research. There are three parameters should be tuned carefully, i.e. D_1 , a frame duration on the right and the left of an energy sample to decide the sample is a maximum energy point or not; Th , a threshold for the ratio of a maximum energy point to a consecutive minimum energy point to decide the point is a local maximum or not; and D_2 , a frame duration to decide a local minimum energy point is syllable boundary or not. Some observations on Indonesian speech dataset produce optimum values for those parameters are $D_1 = 3$, $Th = 1.5$, and $D_2 = 20$.

Figure 2 illustrates the segmentation of an Indonesian utterance “*Dengan skema ini*” (By this scheme) using fuzzy-based smoothing for both global and local normalized STE. In the global normalized STE, two segments /i/ and /ni/ in the utterance produce so low energies that they are flat after fuzzy smoothing and make them are recognized as one syllable /ini/. On the other hand, the fuzzy smoothed local normalized STE gives a better contour for the boundary detection procedure to accurately produces 6 syllables, /de/-/ngan/-/ske/-/ma/-/i/-ni/, as performed by a linguist.

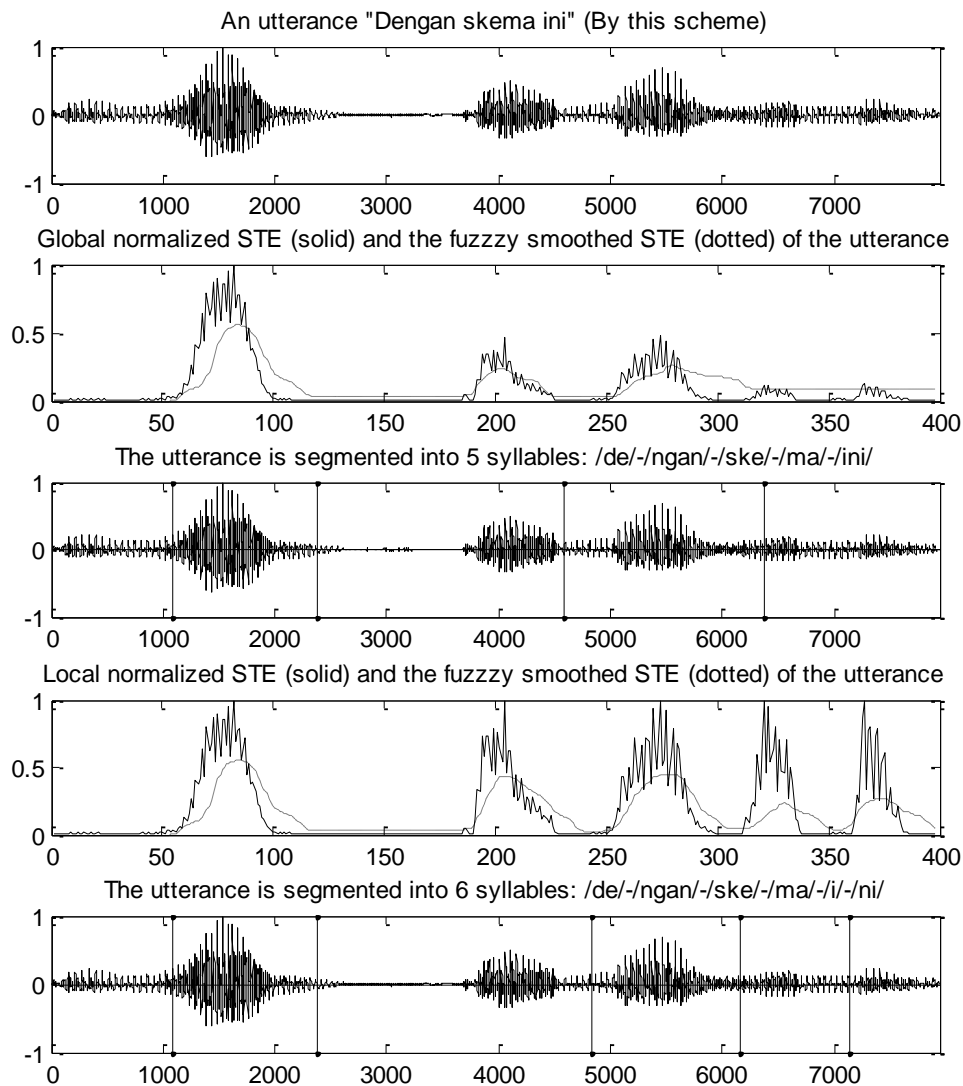


Figure 2 An Indonesian utterance “*Dengan skema ini*” (1), global normalized STE—solid line—and the fuzzy smoothed STE—dotted line (2), the segmentation boundaries produced using global normalized STE (3), local normalized STE and the fuzzy smoothed one (4), the segmentation boundaries produced using local normalized STE (5).

3.6 Postprocessing

Two consecutive Indonesian syllables producing vowel series, i.e. the first syllable ending with a vowel and the second one beginning with a vowel,

commonly have a single energy peak so that they produce deletion error in syllable detection. Hence, a threshold-based splitting procedure is performed to split those such syllables. First, the syllable segments produced by the previous step are scanned and the STE of each segment is recalculated using lower frame size of 9 ms (instead of 10 ms as in [26]) to find more significant variation of energy. A valley in the STE contour can be a syllable boundary if both energy ratio and duration between this valley to its lowest neighbor peak as well as its highest neighbor peak are greater than four predefined thresholds.

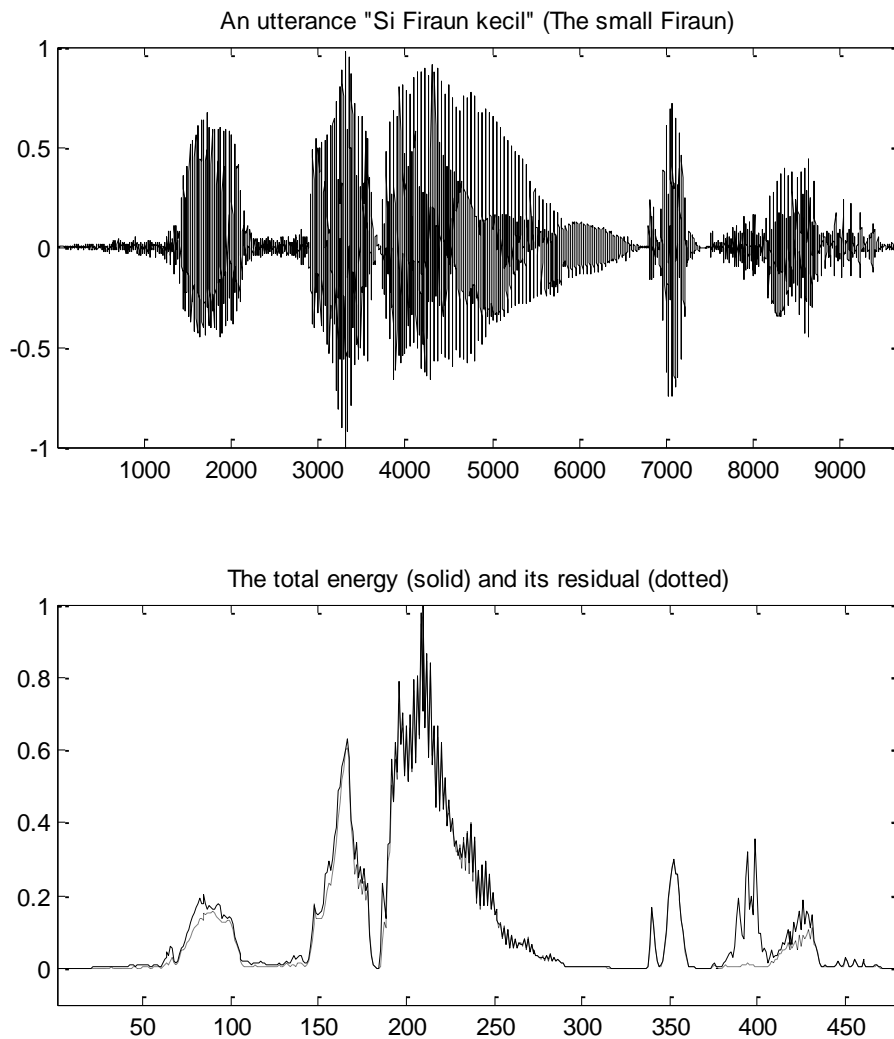


Figure 3 An utterance “*Si Firaun kecil*” (The small Firaun) and the total energy (solid line) as well as its residual energy (dotted line).

As Farsi syllables discussed in [12], Indonesian syllables ending with nonstop nasal consonants such as /m/ and /n/ as well as high energy unvoiced consonants /h/ and glottal stop usually have two energy peaks that cause insertion error. A further observation to fricative consonants /f/, /s/, /z/, /sy/, and /kh/ as in [29] shows that they also cause the such error. Hence, an assimilation procedure as in [26] is adapted to delete the unexpected boundaries. But, in this research, both total and residual energies are calculated using square energy (instead of log root square energy) of the original and the low pass filtered signal respectively, with frame size of 10 ms (instead of 11.6 ms). Here, the signal is filtered with cut-off frequency of 2800 Hz (instead of 1100 Hz). In figure 3, the consonant segments /s/, /f/, and /c/ produce significantly lower residual energies (dotted line) than the total ones (solid line). These facts are exploited to assimilate the such segments into their expected neighbors. Three thresholds to decide the assimilation as described in [26], i.e. *MaxRatio*, *AverageRatio*, and *DecreasingResidualRatio*, are defined empirically by some observations.

4 Result and Discussion

Speech dataset used here is taken from the Indonesian phonetically balanced speech corpus as described in [13] that contains 44,000 utterances from 400 speakers, where each speaker reads 110 sentences. This research takes 220 utterances from two speakers, a male and a female. The dataset of 220 utterances cover varying structures of syllable, from the simple ones such as V, VC, CV, and CVC to the complex ones such as CCVC and CVCCC.

Table 2 The performance of AGN, ALN, ALNS, ALNA, and ALNSA.

Type of ASISS	Accuracy (%)	Insertion (%)	Deletion (%)	Error (%)
AGN	66.26	8.81	20.61	4.32
ALN	82.53	8.49	5.85	3.13
ALNS	86.37	9.99	3.34	0.30
ALNA	81.79	7.51	7.04	3.66
ALNSA	85.57	8.93	4.11	1.39

To see the performance of the proposed additional procedures, five different ASISS systems are developed, i.e. ASISS with global normalization (AGN), ASISS with local normalization (ALN), ALN with splitting (ALNS), ALN with assimilation (ALNA), and ALN with splitting and assimilation (ALNSA). Testing to the 220 Indonesian utterances containing 3,360 syllables gives results listed in table 2, where accuracy is defined as percentage of detected syllables with less than 50 ms boundary error or around 30% of average syllable duration in the dataset. Insertion is percentage of unexpected additional syllable

boundaries occurs in duration of 50 ms. Deletion is percentage of unocurred expected syllable boundaries. Error is percentage of detected syllables with more than 50 ms boundary error.

Compare to the AGN, the proposed ALN gives significantly better performance: improving the accuracy up to 16.27%, reducing the deletion by 14.76%, and slightly reducing the insertion by 0.32%. These results show that the proposed local normalization procedure works very well.

Comparing three other ASISS systems to the ALN gives results as follow: 1) the ALNS reduces deletion by 2.51%, but increases insertion by 1.50%. These results show that the splitting procedure can detect short segments although it oversplits some utterances so that increase the insertion; 2) the ALNA reduces insertion by 0.98%, but increases deletion by 1.19%. This fact indicates that the assimilation procedure does not work very well. It overmerges so many segments that the percentage of deletion increasing more than that of insertion decreasing; 3) the ALNSA slightly reduces deletion by 1.74%, but it increases insertion by 0.44%.

In some cases, the ALNSA can produce the best syllable segments among the others. For instance, see figure 4. A linguist suggests the utterance “*Si Firaun kecil*” (The small Firaun) should be segmented into 5 syllables, /si-/fir-/aun-/ke-/cil/.

The AGN segments the utterance “*Si Firaun kecil*” into 5 syllables, /s-/ifir-/au-/nke-/cil/, shown by solid lines. It produces two insertion errors: boundaries between /s/ and /i/ and between /au/ and /n/, and two deletion errors: syllable boundaries between /si/ and /fir/ and between /aun/ and /ke/. The ALN segments the utterance into 6 syllables, /s-/i-/firaun-/ke-/c-/il/. It produces two insertions, where two single consonant segments /s/ and /c/ should be assimilated into their right segments, and a deletion, i.e. /firaun/ sould be split into /fir/ and /aun/.

The ALN removes one insertion and two deletions produced by the AGN. But, it produces a new insertion, between /c/ and /il/, and a new deletion between /fir/ and /aun/. The ALNS segments the utterance into 7 syllables, /s-/i-/fir-/aun-/ke-/c-/il/. It produces similar segments as those by the ALN, but the deletion between /fir/ and /aun/ can be restored. This result shows that the splitting procedure works accurately. The ALNA segments the utterance into 4 syllables, /si-/firaun-/ke-/cil/. It removes two insertions produced by the ALN. It shows the assimilation procedure works accurately to merge a single consonant segment into the expected neighbor segment. The ALNA can not remove the deletion between /fir/ and /aun/ since the assimilation procedure is

not designed to split any segment. The ALNSA accurately segments the utterance into 5 syllables, i.e. /si-/fir/-/aun/-/ke/-/cil/, with less than 50 ms boundary error, as performed by a linguist. It removes two insertions as well as a deletion produced by the ALN. This result shows that both splitting and assimilation procedures work very well.

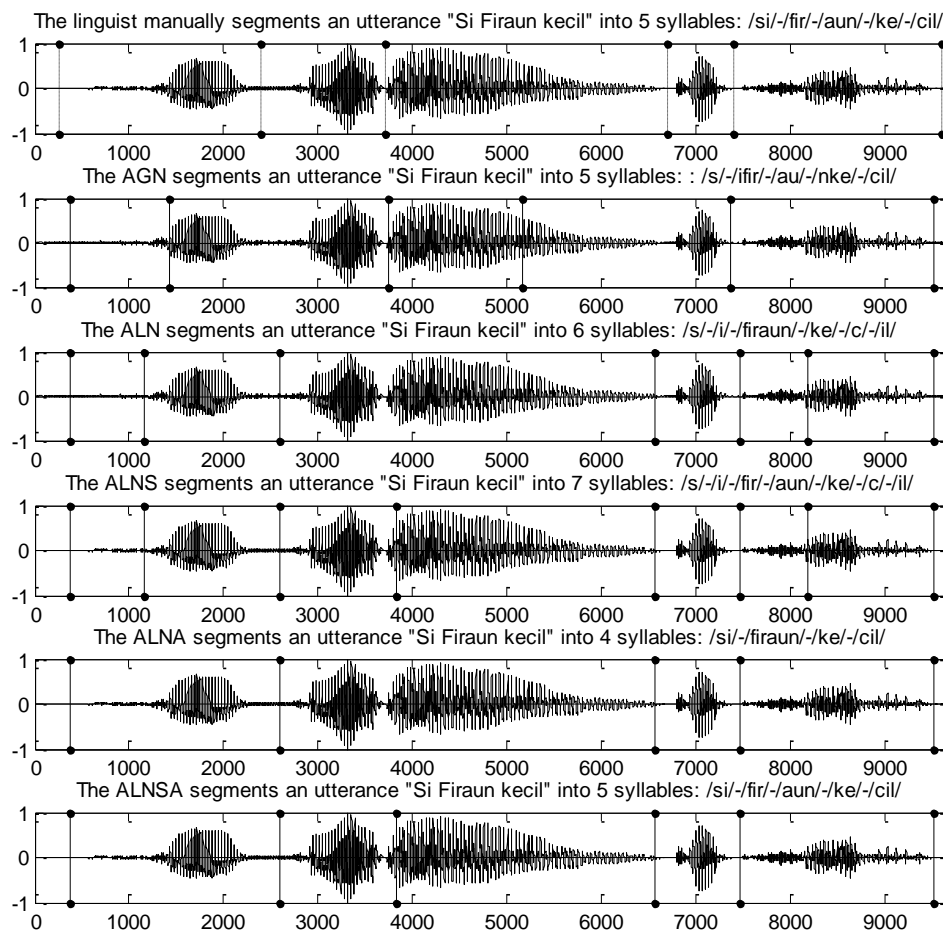


Figure 4 Segmentation of an Indonesian utterance, "Si Firaun kecil" (The small Firaun), by the linguist, AGN, ALN, ALNS, ALNA, and ALNSA.

But, in many other cases, the ALNSA produces worse syllable segments than the ALNS. For instance, see figure 5. An utterance "Kalau dukungan IMF cukup besar terhadap kita" is segmented by a linguist into 22 segments shown

by dotted lines, i.e. /ka/-/lau/-/du/-/ku/-/ngan/-/ik/-/em/-/sp/-/ef/-/sp/-/cu/-/kup/-/be/-/sar/-/sp/-/ter/-/ha/-/dap/-/sp/-/ki/-/sp/-/ta/, where /sp/ is a short pause. The ALN generates 22 segments, shown by solid lines, but it produces two errors: 1) an insertion on the third segment where the segment /lau/ is split into /l/ and /au/; and 2) a deletion on the 18-th segment where two segments, /ha/ and /dap/, are merged as a single segment /hadap/.

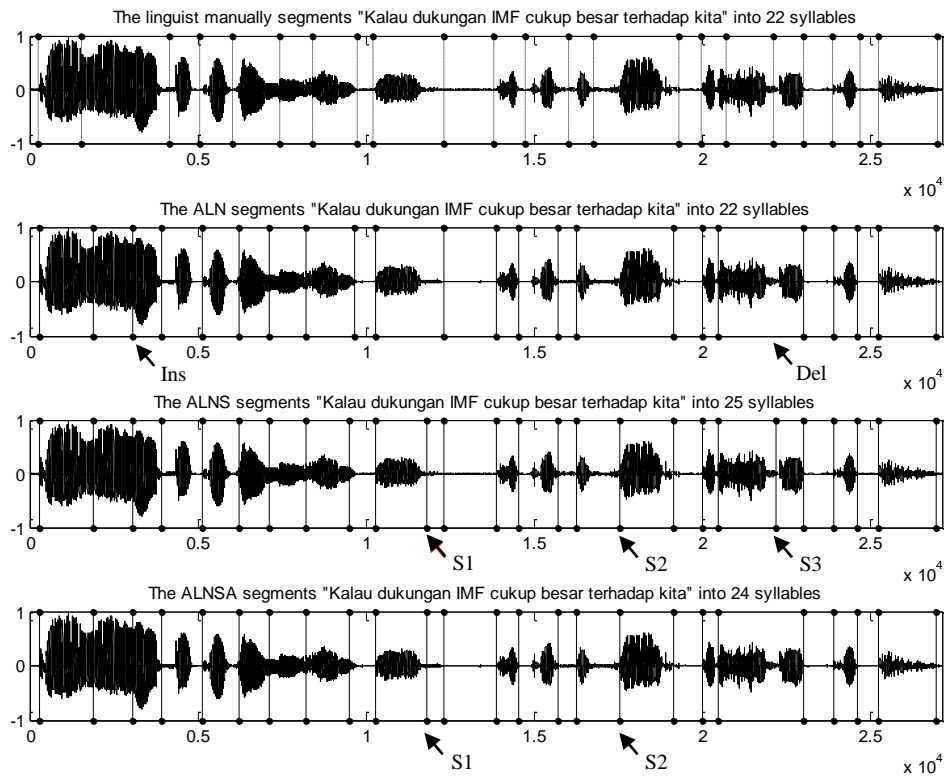


Figure 5 Segmentation of an Indonesian utterance, “Kalau dukungan IMF cukup besar terhadap kita” (If support of IMF is quite big to us), by the linguist, ALN, ALNS, and ALNSA.

The ALNS generates 25 segments shown by solid lines. There are three new additional segments shown by boundaries S1, S2, and S3. It is clear that S1 and S2 are unexpected inserted boundaries, but S3 is an expected boundary that restore the deletion produced by the ALN. These results increase the accuracy of the ALNS with one correct syllable, but they increase the insertion error with two unexpected segments.

The ALNSA unfortunately performs worse than the ALNS by producing 24 segments. It keeps those unexpected inserted boundaries, S1 and S2, but removes the expected boundary S3. The assimilation procedure in the ALNSA does not merge both S1 and S2 into expected previous or next segments, but it assimilates the expected S3 into the previous segment instead. This fact explains why the ALNSA gives lower accuracy than the ALNS as shown in table 2. The weakness of the ALNSA is affected by the assimilation procedure which, in many cases, keeps unexpected insertions and overmerges expected segments.

5 Conclusion

The proposed local normalization significantly improves the performance of fuzzy-based smoothing with global normalization, i.e. increasing the accuracy as well as reducing the insertion and the deletion. Two postprocessing procedures, splitting and assimilation, work in different ways. Splitting missed short syllables sharply reduces the deletion, but slightly increases the insertion. On the other hand, assimilation of a single consonant segment into an expected previous or next segment slightly reduces the insertion, but increases the deletion. Sequential combination of splitting and assimilation unfortunately gives worse accuracy than the splitting procedure only since the assimilation, in many cases, keeps the unexpected insertions and overmerges the expected segments. Hence, an ASISS with local normalization and splitting procedure (ALNS) gives the highest accuracy among the others. In this research, the values for all parameters of the ASISS systems are manually tuned by observations so that they can be not optimum. Hence, an optimization technique, such as evolutionary computation, is possible to be performed to get better optimum values. The assimilation procedure should be improved first before combining sequentially with the splitting procedure.

Acknowledgement

The first author is now a doctoral student in Computer Science Program, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. He is an employee of Telkom Foundation of Education (*Yayasan Pendidikan Telkom*, YPT) as a lecturer at School of Computing, Telkom University (former: Telkom Institute of Technology). This work is supported by YPT with grant number: 15/SDM-06/YPT/2013.

References

- [1] Z. Hu, J. Schalkwyk, E. Barnard, and R. Cole, *Speech Recognition Using Syllable-Like Units*, in Proceedings of ICSLP, pp. 1117–1120 vol.2, 1996.
- [2] M. Jones and P. C. Woodland, *Modelling syllable characteristics to improve a large vocabulary continuous speech recogniser*, in Proceedings of ICSLP, pp. 2171–2714, 1994.
- [3] S. Wu, M. L. Shire, S. Greenberg, and N. Morgan, *Integrating Syllable Boundary Information Into Speech Recognition*, in Proceedings of ICASSP, vol. 2, pp. 987–990, 1997.
- [4] S. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg, *Incorporating information from syllable-length time scales into automatic speech recognition*, in Proceedings of ICASSP, pp. 721–724 vol.2, 1998.
- [5] S. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg, *Performance Improvements Through Combining Phone- and Syllable-Scale Information In Automatic Speech Recognition*, in Proceedings of ICSLP, pp. 459–462, 1998.
- [6] A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchhoff, M. Ordowski, and B. Wheatley, *Syllable-a promising recognition unit for LVCSR*, in Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 207–214, 1997.
- [7] C. D. Bartels and J. A. Bilmes, *Use of syllable nuclei locations to improve ASR*, in Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 335–340, 2007.
- [8] A. Sethy, S. Narayanan, and S. Parthasarthy, *A Syllable Based Approach for Improved Recognition of Spoken Names*, in Proceedings of ISCA Pronunciation Modeling and Lexicon Adaptation, 2002.
- [9] H. Meinedo and J. P. Neto, *The Use of Syllable Segmentation Information in Continuous Speech Recognition Hybrid Systems Applied to The Portuguese Language*, in Proceedings of INTERSPEECH, pp. 927–930, 2000.
- [10] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, *Syllable-based speech recognition using EMG*, in Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2010, pp. 4699–702, 2010.
- [11] R. Janakiraman, J. C. Kumar, and H. A. Murthy, *Robust syllable segmentation and its application to syllable-centric continuous speech*

- recognition*, in Proceedings of National Conference on Communications (NCC), pp. 1–5, 2010.
- [12] G. Sheikhi and F. Almasganj, *Segmentation of Speech into Syllable Units using Fuzzy Smoothed Short Term Energy Contour*, in Proceedings of The 18th Iranian Conference on BioMedical Engineering, no. December, pp. 195–198, 2011.
- [13] Suyanto and J. Adityatama, *Yooi: An Indonesian Short Message Dictation*, International Journal of Intelligent Information Processing (IJIP), vol. 3, no. 4, pp. 68–74, 2012.
- [14] Suyanto and S. Hartati, *Design of Indonesian LVCSR Using Combined Phoneme and Syllable Models*, in Proceedings of The 7th International Conference on Information & Communication Technology and Systems (ICTS), pp. 191–196, 2013.
- [15] P. Mermelstein, *Automatic segmentation of speech into syllabic units*, J. Acoust. Soc. Am., vol. 58, no. 4, pp. 880–883, 1975.
- [16] E. A. Kaur and E. T. Singh, *Segmentation of Continuous Punjabi Speech Signal into Syllables*, in Proceedings of The World Congress on Engineering and Computer Science (WCECS), vol. I, pp. 20–23, 2010.
- [17] E. Lewis, *Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis*, in Proceedings of EUROSPEECH, pp. 1–5, 2001.
- [18] T. Nagarajan, H. A. Murthy, and R. M. Hegde, *Segmentation of speech into syllable-like units*, in Proceedings of EUROSPEECH, pp. 2893–2896, 2003.
- [19] K. V. Prasad, T. Nagarajan, and H. A. Murthy, *Automatic segmentation of continuous speech using minimum phase group delay functions*, Speech Communication, vol. 42, no. 3–4, pp. 429–446, Apr. 2004.
- [20] I. Kopeček, *Speech Recognition and Syllable Segments*, in Proceedings of The 2nd International Workshop (TSD), pp. 203–208, 1999.
- [21] S. Nakagawa and Y. Hashimoto, *A method for continuous speech segmentation using HMM*, in Proceedings of The 9th International Conference on Pattern Recognition, pp. 960 – 962 vol.2, 1988.
- [22] H. A. Murthy and B. Yegnanarayana, *Group delay functions and its applications in speech technology*, Sadhana, vol. 36, no. 5, pp. 745–782, 2011.
- [23] T. Jianhua and H. U. Hain, *Syllable Boundaries based Speech Segmentation in Demi-Syllable Level for Mandarin with HTK*, in Proceedings of Oriental COCOSDA, 2002.

- [24] L. Shastri, S. Chang, and S. Greenberg, *Syllable Detection And Segmentation Using Temporal Flow Neural Networks*, in Proceedings of The 14th International Congress of Phonetic Sciences, 1999.
- [25] H. R. Ptzinger, S. Burger, and S. Heid, *Syllable detection in read and spontaneous speech*, in Proceedings of ICSLP, pp. 1261–1264 vol.2, 1996.
- [26] S. Fische and N. Federico, *A syllable segmentation algorithm for English and Italian*, in Proceedings of EUROSPEECH, pp. 2913–2916, 2003.
- [27] R. Villing, J. Timoney, T. Ward, and J. Costello, *Automatic Blind Syllable Segmentation for Continuous Speech*, in Proceedings of ISSC, 2004.
- [28] P. Santiprabhob, J. Chaiareerat, and R. Cheirsilp, *A Framework for Connected Speech Recognition for Thai Language*, AU J.T., vol. 8, no. 3, pp. 113–123, 2005.
- [29] Alwi, H., Dardjowidjojo, S., Lapoliwa, H., and Moeliono, A.M., *Tata bahasa baku bahasa Indonesia (The standart Indonesian grammar)*, Jakarta, Balai Pustaka, 1998.