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

Suyanto[1] & Agfianto Eko Putra[2]

[1]School of Computing, Telkom University
Jalan Telekomunikasi Terusan Buah Batu, Bandung 40257, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Gadjah Mada University
Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia
Email: suy@ittelkom.ac.id

Abstract. This paper discusses the usage of the short-term energy contour of speech smoothed by a fuzzy-based method to automatically segment it into syllabic units. Two new additional procedures, local normalization and postprocessing, are proposed to adapt to the Indonesian language. Testing to 220 Indonesian utterances showed that the local normalization significantly improved the performance of the fuzzy-based smoothing. In the postprocessing procedure, splitting and assimilation work in different ways. The splitting of missed short syllables sharply reduced deletion, but slightly increased insertion. On the other hand, the assimilation of a single consonant segment into an expected previous or next segment slightly reduced insertion, but increased deletion. The use of splitting gave a higher accuracy than the assimilation and 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 on syllabic units can be used to improve the performance of flat start-based automatic speech recognition (ASR) [1]-[11]. In 2010, Janakiraman et al. [11] reported that incorporating information on syllable boundaries into English ASR reduced both computational complexity and word error rate (WER) significantly compared to 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.

Every language has unique characteristics. For example, English and Indonesian have different syllable patterns. A study of telephone conversations and switchboard corpus by Su-Lin Wu [3] has shown that English has 80% monosyllabic words and 85% of them are simple structures (V, VC, CV, CVC)

Received October 1st, 2013, 1st Revision April 2nd, 2014, 2nd Revision May 26th, 2014, Accepted for publication May 30th, 2014.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2014.8.2.2

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

This research focuses on syllable segmentation for the Indonesian language. A segmentation method in [12] that was designed for the Farsi language, with simple syllable structures CV(C)(C), was adapted and tested to the Indonesian speech dataset of the clean speech corpus as described in [13]. Some modifications and 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 related works about syllable segmentation for some languages, Section 3 describes the proposed Indonesian syllable segmentation, Section 4 reports experimental results 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) the time domain, as in [12], [15]-[17]; 2) the frequency domain, as in [11], [18]-[25]; and a combination of both, as in [26]-[28]. The time domain approaches mostly use the short-term energy (STE) contour smoothed by a smoothing algorithm, while the frequency domain approaches exploit cepstrum features.

The time domain approach in [16] simply uses a plain STE contour and a threshold to detect the locations of the start and end of a syllable. This method works very well but only for short-sentence utterances. In [12], the plain STE contour was first smoothed by fuzzy-based smoothing before defining the syllable boundaries using a threshold-based method. The usage of fuzzy-based smoothing gave a much higher accuracy, 93.8% for a Farsi speech dataset, than the common moving average smoothing method. Unfortunately, this method produced a 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 a minimum phase group delay function [18],[19]. This method can be 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].

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

3 The Proposed Syllable Segmentation

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

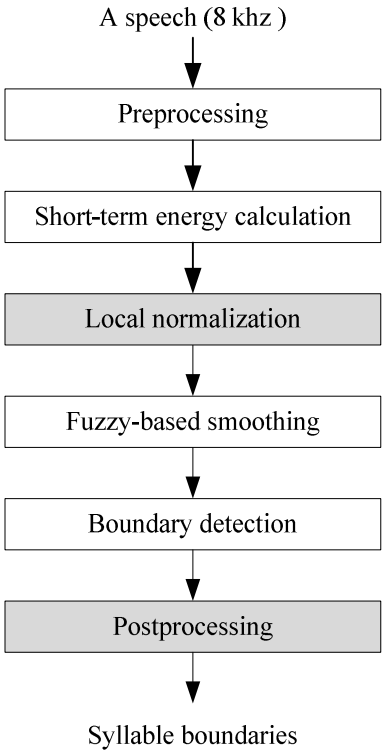

Figure 1 Block diagram of ASISS.

3.1 Preprocessing

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

$$y_i = x_i - \alpha x_{i-1} \tag{1}$$

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

Long sentence speech commonly contains a number of silences that are so long that it is quite hard to find the accurate syllable boundaries in such speech. Hence, a threshold-based procedure of silence removal is performed using both energy and duration thresholds on the STE. Thus, 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

The STE contour can be produced using different formulas, such as absolute, square, root mean square, Teager, and modified Teager. Sheikhi and Almasganj have shown that the Teager energy gives the best accuracy for the Farsi speech dataset [12]. However, in this research, the square energy as in Eq. (2) is used, because it can increase the difference between a low signal energy and a higher one and it empirically gives the best accuracy for the Indonesian speech dataset.

$$E = \sum_{i=1}^{N} S_i^2 \tag{2}$$

3.3 Local Normalization

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

3.4 Fuzzy-Based Smoothing

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

The membership functions for any rule and the term ***most*** in the fuzzy linguistic rules as well as the activity degree of any fuzzy group were adapted from [12]. The membership functions for all fuzzy rules are described by Eq. (3), where $x_i = E_i - \hat{E}_i$, i.e. crisp valued inputs come from the speech energy substracted 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 $w = 0.18$ was found through a number of 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 positive **then** output is very small positive
2	**if** most inputs are small positive **then** output is small positive
3	**if** most inputs are medium positive **then** output is medium positive
4	**if** most inputs are big positive **then** output is big positive
5	**if** most inputs are very big positive **then** output is very big positive
6	**if** most inputs are very small negative **then** output is very small negative
7	**if** most inputs are small negative **then** output is small negative
8	**if** most inputs are medium negative **then** output is medium negative
9	**if** most inputs are big negative **then** output is big negative
10	**if** most inputs are very big negative **then** output is very big negative
11	**else** output is zero

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

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

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

The activity degree of any fuzzy group is described by Eq. (5) [12].

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

Output of the fuzzy-based smoothing is a correlation product in Eq. (6).

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

Finally, the fuzzy smoothed energy is calculated by Eq. (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] was adapted in this research. There are three parameters that should be tuned carefully, i.e. D_1, the frame duration on the right and the left of an energy sample to decide if the sample is a maximum energy point or not; *Th*, the threshold for the ratio of a maximum energy point to a consecutive minimum energy point to decide if the point is a local maximum or not; and D_2, the frame duration to decide if a local minimum energy point is syllable boundary or not. Observation of the Indonesian speech dataset produced the following optimum values for those parameters: $D_1 = 3$, $Th = 1.5$, and $D_2 = 20$.

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

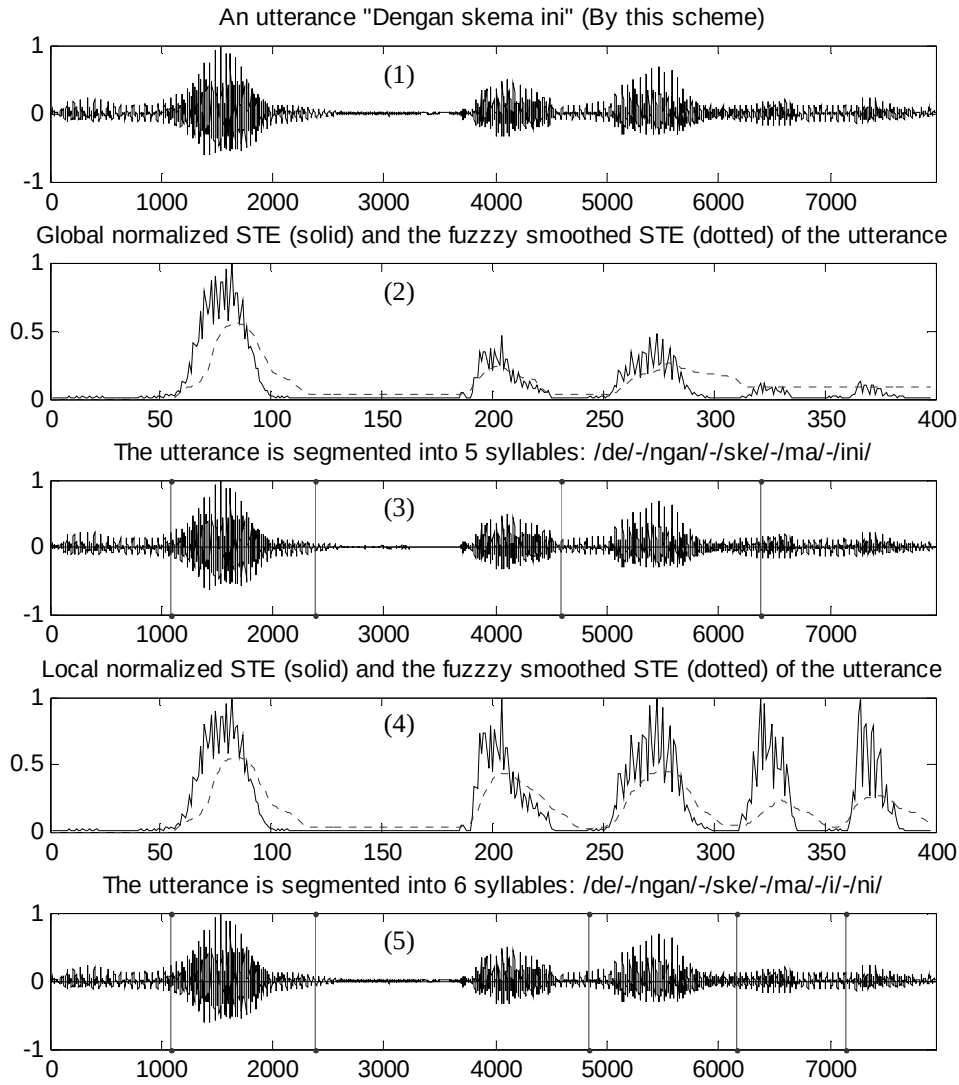

Figure 2 Indonesian utterance "*Dengan skema ini*" (1); global normalized STE (solid line) and fuzzy smoothed STE (dotted line) (2); segmentation boundaries produced using global normalized STE (3); local normalized STE and fuzzy smoothed STE (4); segmentation boundaries produced using local normalized 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 begining with a vowel,

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

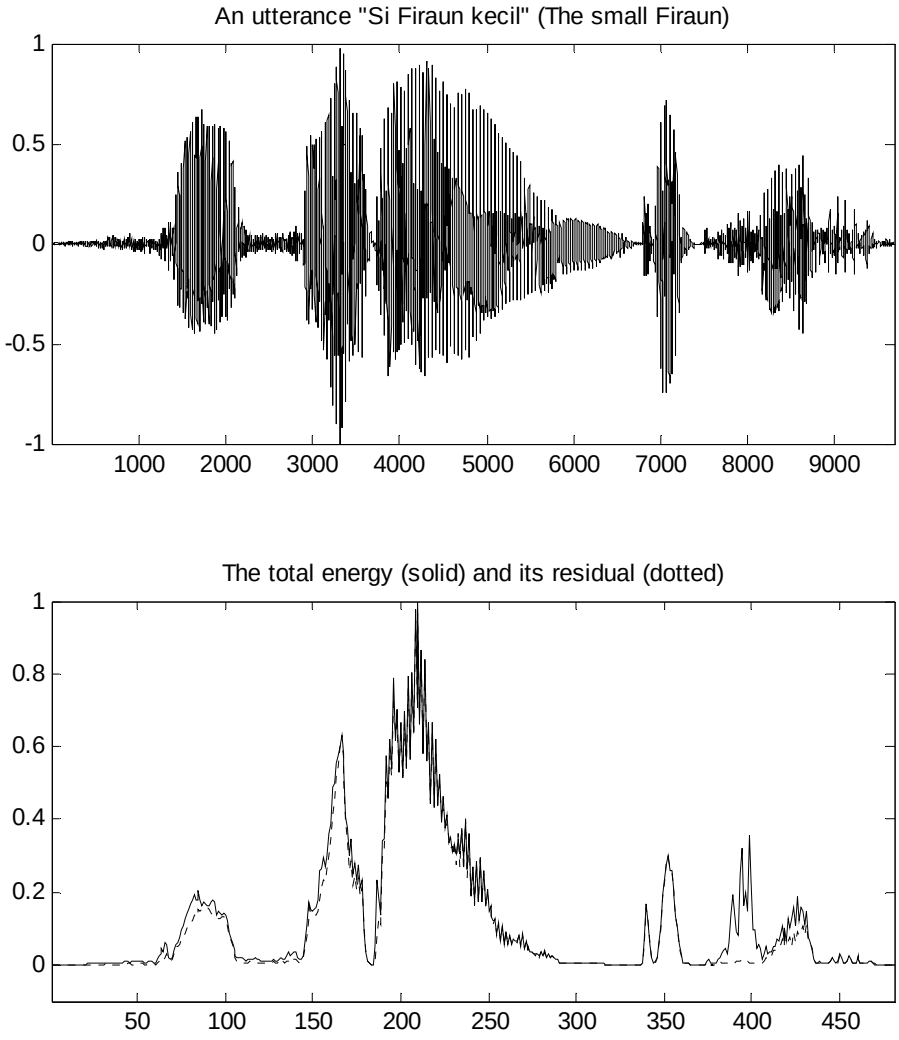

Figure 3 The utterance "*Si Firaun kecil*" (The small Firaun) and the total energy (solid line) as well as its residual energy (dotted line).

Like similar 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 can cause insertion errors. Further observation of the fricative consonants /f/, /s/, /z/, /sy/, and /kh/ as discussed in [29], shows that they also cause such errors. Hence, the assimilation procedure from [26] was adapted to delete the unexpected boundaries. However, in this research, both the total and residual energies are calculated using the square energy (instead of the log root square energy) of the original and the low pass filtered signal respectively, with a frame size of 10 ms (instead of 11.6 ms). Here, the signal is filtered with a cut-off frequency of 2800 Hz (instead of 1100 Hz). Figure 3 shows that the consonant segments /s/, /f/, and /c/ produce significantly lower residual energies (dotted line) than the total energies (solid line). This fact is exploited to assimilate such segments into their expected neighbors. The three thresholds to decide assimilation, as described in [26], i.e. *MaxRatio*, *AverageRatio*, and *DecreasingResidualRatio*, were defined empirically on the basis of several observations.

4 Results and Discussion

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

Table 2 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

In order to see the performance of the proposed additional procedures, five different ASISS systems were 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 gave the results listed in Table 2, where accuracy is defined as the percentage of detected syllables with boundary errors smaller than 50 ms or around 30% of the average syllable duration in the dataset. Insertion is the percentage of unexpected additional syllable boundaries that occurred within a duration of 50

ms. Deletion is the percentage of unoccured expected syllable boundaries. Error is the percentage of detected syllables with boundary errors larger than 50 ms.

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

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

In some cases, ALNSA produced the best syllable segmenting. See for instance figure 4. A linguist would suggest that the utterance "*Si Firaun kecil*" (The small Firaun) should be segmented into 5 syllables, /si/-/fir/-/aun/-/ke/-/cil/.

AGN segmented the utterance "*Si Firaun kecil*" into 5 syllables, /s/-/ifir/-/au/-/nke/-/cil/, as indicated by the solid lines. It produced 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/.

ALN segmented the utterance into 6 syllables, /s/-/i/-/firaun/-/ke/-/c/-/il/. It produced two insertions, where two 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 removed one insertion and two deletions produced by AGN. However, it produced a new insertion, between /c/ and /il/, and a new deletion between /fir/ and /aun/.

ALNS segmented the utterance into 7 syllables, /s/-/i/-/fir/-/aun/-/ke/-/c/-/il/. It produced similar segments as those produced by ALN, but the deletion between /fir/ and /aun/ would be restored later. This result shows that the splitting procedure works accurately.

ALNA segmented the utterance into 4 syllables, /si/-/firaun/-/ke/-/cil/. It removed two insertions produced by ALN. This shows that the assimilation procedure works accurately in merging a single consonant segment with the

expected neighbor segment. ALNA could not remove the deletion between /fir/ and /aun/ since the assimilation procedure is not designed to split any segments.

ALNSA accurately segmented the utterance into 5 syllables, i.e. /si/-/fir/-/aun/-/ke/-/cil/, with a boundary error smaller than 50 ms, the same segmentation as performed by a linguist. It removed two insertions as well as a deletion produced by ALN. This result shows that both the splitting and the assimilation procedures work very well.

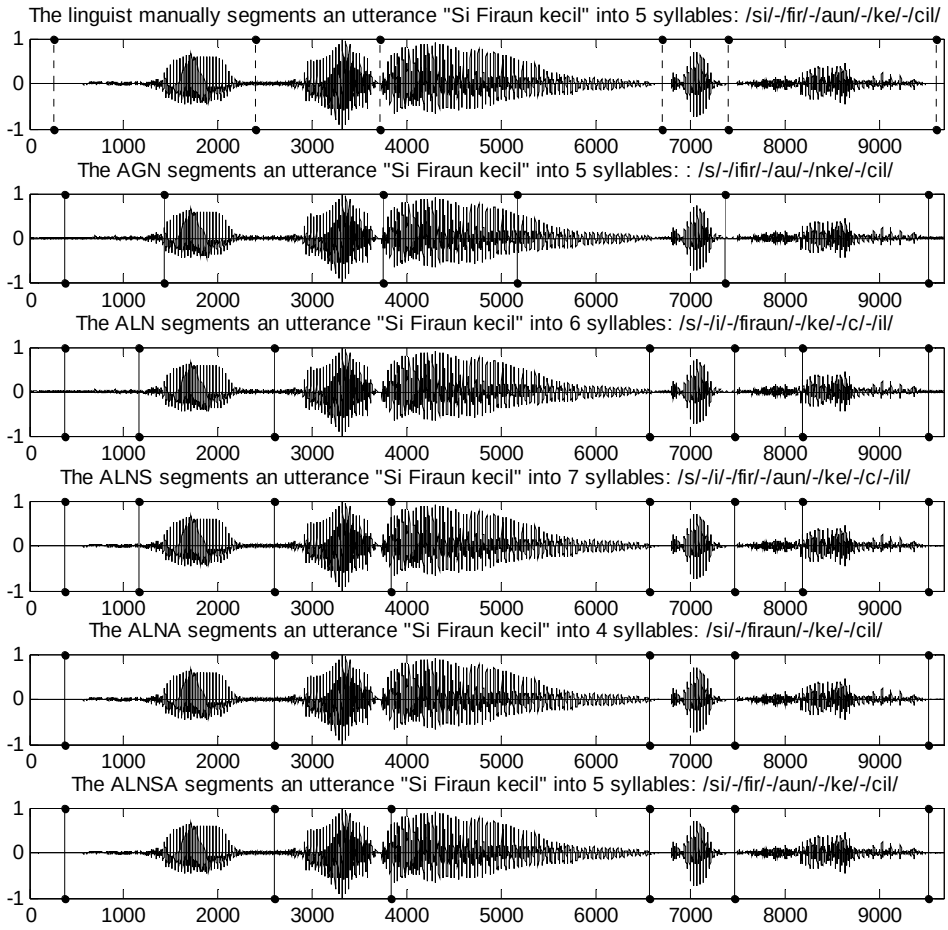

Figure 4 Segmentation of an Indonesian utterance, "*Si Firaun kecil*" (The small Firaun), by a linguist, AGN, ALN, ALNS, ALNA, and ALNSA.

However, in many other cases, ALNSA produced worse syllable segmenting than ALNS. See for instance Figure 5. The utterance "*Kalau dukungan IMF cukup besar terhadap kita*" is segmented by a linguist into 22 segments as indicated by the 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. ALN generated 22 segments, indicated by the solid lines, but it produced two errors: 1) an insertion in the third segment, where the segment /lau/ is split into /l/ and /au/; and 2) a deletion in the 18-th segment, where two segments, /ha/ and /dap/, are merged as a single segment /hadap/.

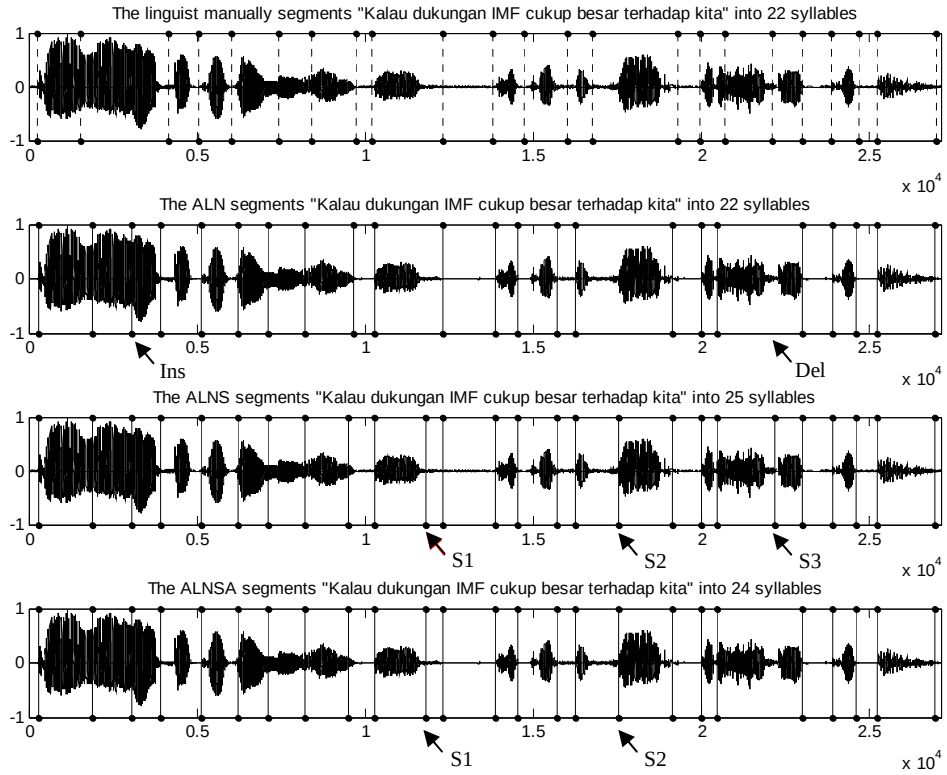

Figure 5 Segmentation of an Indonesian utterance, "Kalau dukungan IMF cukup besar terhadap kita" (If support from the IMF is quite big to us), by a linguist, ALN, ALNS, and ALNSA.

ALNS generated 25 segments as indicated by the solid lines. There were 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 restores the deletion produced by ALN. These results increased the

accuracy of ALNS with one correct syllable, but they also increased the number of insertion errors with two unexpected segments.

ALNSA unfortunately performed worse than ALNS by producing 24 segments. It kept the unexpected inserted boundaries S1 and S2, but removed the expected boundary S3. The assimilation procedure in ALNSA did not merge both S1 and S2 into expected previous or next segments, but assimilated the expected S3 into the previous segment instead. This explains why ALNSA gave a lower accuracy than ALNS, as shown in Table 2. The weakness of ALNSA is affected by the assimilation procedure, which, in many cases, keeps unexpected insertions and overmerges expected segments.

5 Conclusions

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

Acknowledgements

The first author is now a doctoral student in the 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 the 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] Hu, Z., Schalkwyk, J., Barnard, E. & Cole, R., *Speech Recognition Using Syllable-Like Units*, in Proceedings of The 4th ICSLP, Philadelphia, PA, USA, International Speech Communication Association (ISCA), **2**, pp. 1117-1120, 1996.

[2] Jones, M. & Woodland, P.C., *Modelling Syllable Characteristics to Improve a Large Vocabulary Continuous Speech Recogniser*, in Proceedings of The 3rd ICSLP, Yokohama, Japan, International Speech Communication Association (ISCA), pp. 2171-2714, 1994.

[3] Wu, S., Shire, M. L., Greenberg, S. & Morgan, N., *Integrating Syllable Boundary Information Into Speech Recognition*, in Proceedings of ICASSP, Munich, Bavaria, Germany, IEEE Signal Processing Society, **2**, pp. 987-990, 1997.

[4] Wu, S., Kingsbury, B.E.D., Morgan, N. & Greenberg, S., *Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition*, in Proceedings of ICASSP, Seattle, Washington, USA, IEEE Signal Processing Society, **2**, pp. 721-724, 1998.

[5] Wu, S., Kingsbury, B.E.D., Morgan, N. & Greenberg, S., *Performance Improvements Through Combining Phone- and Syllable-Scale Information In Automatic Speech Recognition*, in Proceedings of The 5th ICSLP, Sydney, Australia, International Speech Communication Association (ISCA), pp. 459-462, 1998.

[6] Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G.,. Kirchhoff, K., Ordowski, M. & Wheatley, B., *Syllable-a Promising Recognition Unit for LVCSR*, in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Santa Barbara, CA, IEEE Signal Processing Society, pp. 207-214, 1997.

[7] Bartels, C.D. & Bilmes, J.A., *Use of Syllable Nuclei Locations to Improve ASR*, in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan, IEEE Signal Processing Society, pp. 335-340, 2007.

[8] Sethy, A., Narayanan, S. & Parthasarthy, S., *A Syllable Based Approach for Improved Recognition of Spoken Names*, in Proceedings of The ISCA Pronunciation Modeling and Lexicon Adaptation, Aspen Lodge, Estes Park, Colorado, USA, International Speech Communication Association (ISCA), pp. 30-35, 2002.

[9] Meinedo, H. & Neto, J.P., *The Use of Syllable Segmentation Information in Continuous Speech Recognition Hybrid Systems Applied to The Portuguese Language*, in Proceedings of INTERSPEECH, Makuhari, Chiba, Japan, International Speech Communication Association (ISCA), pp. 927-930, 2000.

[10] Lopez-Larraz, E., Mozos O.M., Antelis, J.M. & Minguez, J., *Syllable-Based Speech Recognition Using EMG*, in Proceedings of The 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, IEEE, **2010**, pp. 4699-4702, 2010.

[11] Janakiraman, R., Kumar, J.C. &. Murthy, H.A., *Robust Syllable Segmentation and Its Application to Syllable-Centric Continuous Speech Recognition*, in Proceedings of The 16th National Conference on Communications (NCC), Chennai, India, Joint Telematics Group of IITs & IISc, pp. 1-5, 2010.

[12] Sheikhi, G. & Almasganj, F., *Segmentation of Speech into Syllable Units using Fuzzy Smoothed Short Term Energy Contour*, in Proceedings of The 18th Iranian Conference on BioMedical Engineering, Tehran, Iran, IEEE Iran Section, pp. 195-198, 2011.

[13] Suyanto & Adityatama, J., *Yooi: An Indonesian Short Message Dictation*, International Journal of Intelligent Information Processing, **3**(4), pp. 68-74, 2012.

[14] 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, Informatics Department, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember (ITS) Surabaya, pp. 191-196, 2013.

[15] Mermelstein, P., *Automatic Segmentation of Speech Into Syllabic Units*, J. Acoust. Soc. Am., **58**(4), pp. 880-883, 1975.

[16] Kaur E.A. & Singh, E.T., *Segmentation of Continuous Punjabi Speech Signal into Syllables*, in Proceedings of World Congress on Engineering and Computer Science (WCECS), San Francisco, USA, The International Association of Engineers (IAENG), **I**, pp. 20-23, 2010.

[17] Lewis, E., *Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis*, in Proceedings of EUROSPEECH, Aalborg, Denmark, International Speech Communication Association (ISCA), pp. 1-5, 2001.

[18] Nagarajan, T., Murthy, H.A. & Hegde, R.M. *Segmentation of Speech Into Syllable-Like Units*, in Proceedings of EUROSPEECH, pp. 2893-2896, 2003.

[19] Prasad, K.V., Nagarajan, T. & Murthy, H.A., *Automatic Segmentation of Continuous Speech Using Minimum Phase Group Delay Functions*, Speech Communication, **42**(3-4), pp. 429-446, Apr. 2004.

[20] Kopeček, I., *Speech Recognition and Syllable Segments*, in Proceedings of the 2nd International Workshop (TSD), Plzen, Czech Republic, The Faculty of Applied Sciences, University of West Bohemia, Plzen (Pilsen) and the Faculty of Informatics, Masaryk University, Brno, pp. 203-208, 1999.

[21] Nakagawa, S. & Hashimoto, Y., *A Method for Continuous Speech Segmentation Using HMM*, in Proceedings of The 9th International Conference on Pattern Recognition, Rome, Italy, IEEE Computer Society and International Association for Pattern Recognition, **2**, pp. 960-962, 1988.

[22] Murthy, H.A. & Yegnanarayana, B., *Group Delay Functions and Its Applications in Speech Technology*, Sadhana, **36**(5), pp. 745-782, 2011.

[23] Jianhua, T. & Hain, H.U., *Syllable Boundaries based Speech Segmentation in Demi-Syllable Level for Mandarin with HTK*, in Proceedings of International joint conference of SNLP-Oriental COCOSDA, Hua Hin, Prachuapkirikhan, Thailand, Thanaruk Theeramunkong and Virach Sornlertlamvanich, 2002.

[24] Shastri, L., Chang, S. & Greenberg, S., *Syllable Detection and Segmentation Using Temporal Flow Neural Networks*, in Proceedings of The 14th International Congress of Phonetic Sciences, San Francisco, USA, University of California, pp. 1721-1724, 1999.

[25] Ptzinger, H.R., Burger, S., & Heid, S., *Syllable Detection in Read and Spontaneous Speech*, in Proceedings of ICSLP, **2**, pp. 1261-1264, 1996.

[26] Fisiche, S. & Federico, N., *A Syllable Segmentation Algorithm for English and Italian*, in Proceedings of INTERSPEECH, Geneva, Switzerland, International Speech Communication Association (ISCA), pp. 2913-2916, 2003.

[27] Villing, R., Timoney, J., Ward, T. & Costello, J., *Automatic Blind Syllable Segmentation for Continuous Speech*, in Proceedings of the Irish Signals and Systems Conference (ISSC), Queen's University Belfast, Northern Ireland, Institution of Electrical Engineers, 2004.

[28] Santiprabhob, P., Chaiareerat, J. & Cheirsilp, R., *A Framework for Connected Speech Recognition for Thai Language*, AU Journal of Technology, **8**(3), pp. 113-123, 2005.

[29] Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, A.M., *The Standard Indonesian Grammar (Tata Bahasa Baku Bahasa Indonesia)*, 3rd Edition, Jakarta, Balai Pustaka, 1998.