Thai Phoneme Segmentation using Dual-Band Energy Contour

S. Ratsameewichai, N. Theera-Umpon, J. Vilasdechanon, S. Uatrongjit, and K. Likit-Anurucks
Electronic Circuits and Systems Research Laboratory,
Department of Electrical Engineering, Chiang Mai University, Chiangmai 50200, Thailand
E-mail: sermsak@eng.cmu.ac.th

Abstract: In this paper, a new technique for Thai isolated speech phoneme segmentation is proposed. Based on Thai speech feature, the isolated speech is first divided into low and high frequency components by using the technique of wavelet decomposition. Then the energy contour of each decomposed signal is computed and employed to locate phoneme boundary. To verity the proposed scheme, some experiments have been performed using 1,000 syllables data recorded from 10 speakers. The accuracy rates are 96.0, 89.9, 92.7 and 98.9% for initial consonant, vowel, final consonant and silence, respectively.

1. Introduction

Speech recognition has played significant roles in many applications such as telecommunications, military, medicine, etc [1–2, 4–5, 7–9]. Speech recognition using whole-word speech models needs sufficiently large, sometimes impractically large, training set. Therefore, a more efficient speech representation such as subword speech units is essential for large vocabulary systems [7].

In phoneme based speech recognition, the recognition rate highly depends on the phoneme segmentation quality [3]. In general, phoneme segmentation is performed using the signal energy contour [9] and the pitch period [4]. However, some consonants and vowels produce similar energy curves and long vowel pronunciation sometimes causes an abrupt energy change. These effects introduce difficulties in phoneme segmentation.

In Thai speech, it is known that the initial consonant usually contains either low or high frequency components. While the vowel has both low and high frequency components, much of its energy is contained in the low frequency band. It is also known that the final consonant consists of only low frequency components. These observations suggest that the phoneme borders may be located by considering the energy contour of both low and high frequency components separately.

Based on these features, a new technique for Thai isolated speech phoneme segmentation is proposed. In our method, an isolated speech is first divided into low and high frequency components by using wavelet decomposition technique [6]. Then the energy contour of each decomposed signal is computed and employed to locate phoneme boundary.

This paper is organized as follows. First we briefly describe the decomposition of speech signal into low and high frequency components using wavelet. Then the proposed phoneme segmentation method is explained. Experimental results are given in section 4. Finally some conclusions are discussed.

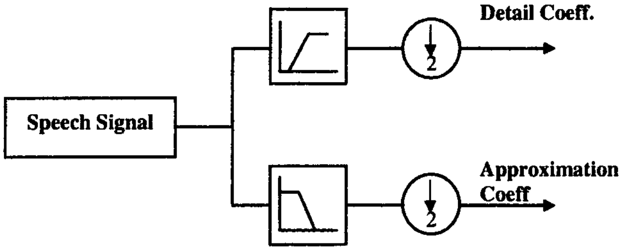

Figure 1. One level wavelet decomposition.

2. Speech Signal Decomposition

In the proposed segmentation method, it is desired to decompose the speech signal into low and high frequency components. It is possible to design special low-pass and high-pass filters for this task. However, in this paper, the technique of wavelet decomposition is applied. In Figure 1, the block diagram of one level wavelet decomposition is presented. After decomposing, we obtain two sets of data namely the approximation and the detail coefficients. From the theory of wavelet, it is known that the approximate coefficients contain low frequency signal component while the detail coefficients contain high frequency signal component. Hence, we can reconstruct the low frequency component $A[n]$ by setting detail coefficients to zero and perform the wavelet reconstruction. Similarly, the high frequency component $D[n]$ is reconstructed by setting approximation coefficients to zero and perform wavelet reconstruction as outlined in Figure 2.

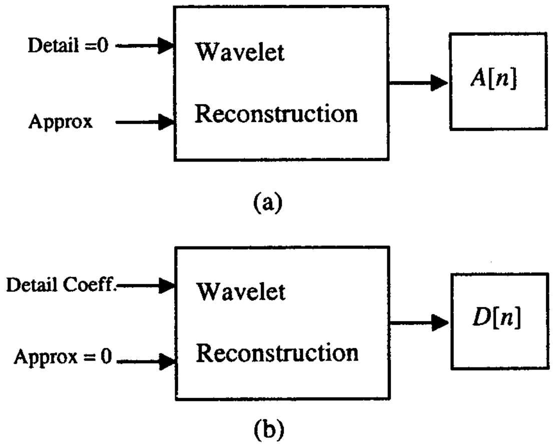

Figure 2. (a) Low and (b) High frequency components signal decomposition.

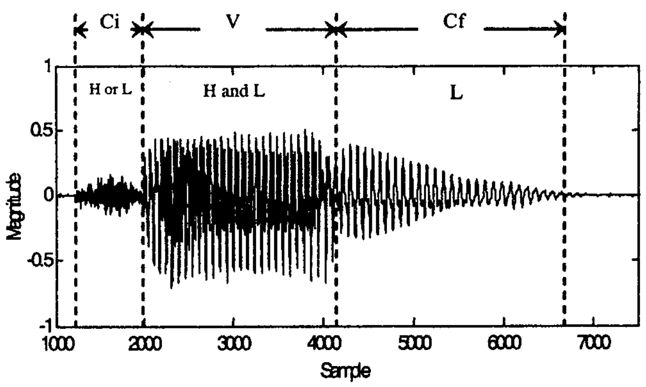

Figure 3. Feature of Thai speech.

3. Phoneme Segmentation

An example of Thai speech is shown in Figure 3. It shows that the signal can be divided into three parts namely the initial consonant Ci, the vowel V, and the final consonant Cf. Time variation of signal's frequency in the V part determines the tone. In Thai speech, the Ci part contains either high or low frequency signal depending on the consonant class. Hence if Ci part contains high frequency component, then its high frequency component energy contour is used to locate the start position of syllable. The energy contour of its low frequency component is used to locate the start position of the vowel. On the other hand, if Ci part contains low frequency component, then we use the energy content of the low frequency part to locate the start position of the syllable. Its high frequency component energy contour is used to locate the start position of the vowel. The stop position of the vowel is determined by considering the energy contour of high frequency component. The stop position of the syllable is determined by using the energy contour of low frequency component of the signal.

3.1 Algorithm

The proposed segmentation technique is performed as shown below. Concurrently with the algorithm description, we provide an example of a word "team" [ti:m] and corresponding results in Figure 4. Figure 4(a) depicts the signal of the input word.

1. The signal is decomposed by using the method described in section 2 into the approximation ($A_1[n]$) and the detail ($D_1[n]$) signals. The resulting $A_1[n]$ and $D_1[n]$ are shown in Figures 4(b) and (c), respectively. We can consider $A_1[n]$ as low frequency component

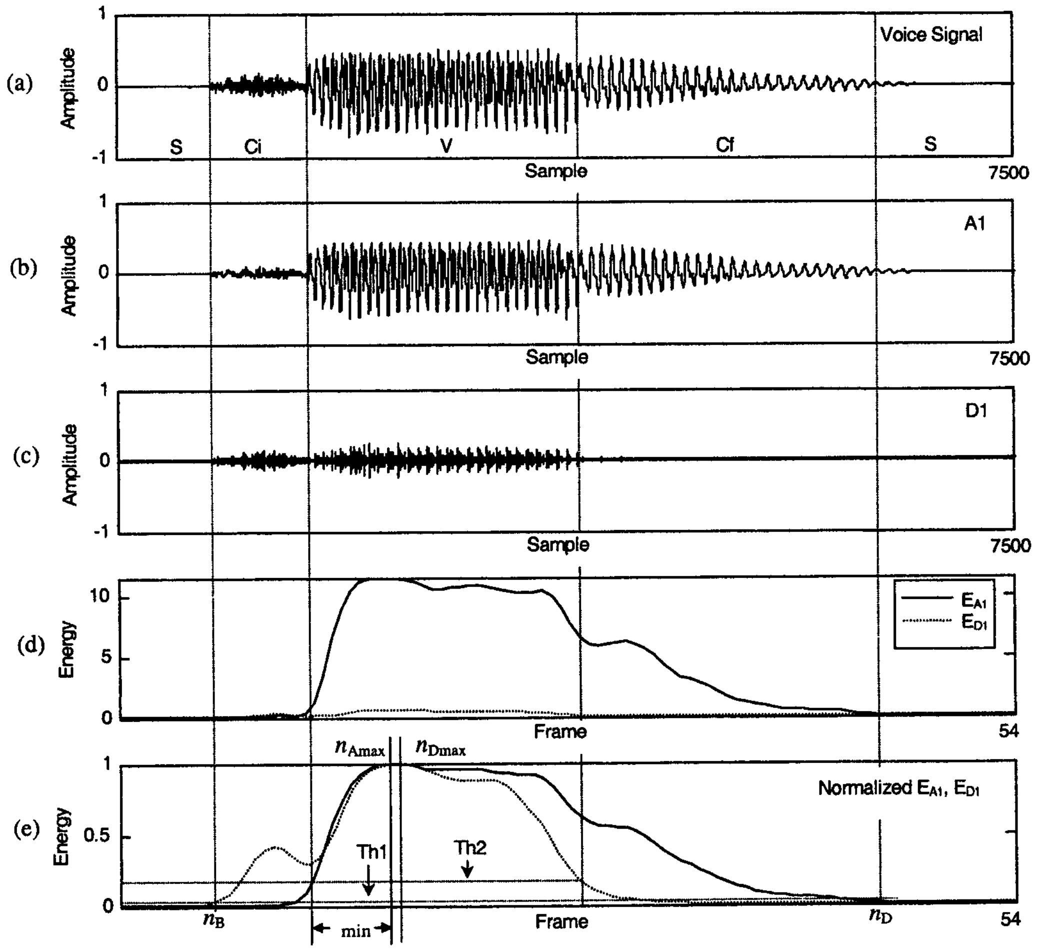

Figure 4. Example of a sound team [ti:m] and phoneme segmentation.

and $D_1[n]$ as high frequency component of the voice signal.

2. Compute $E_{A1}[n]$ and $E_{D1}[n]$, i.e., the normalized square energy contours of $A_1[n]$ and $D_1[n]$, respectively. The resulting energy contours are shown in figure 4(e). Let $n_{A\max}$ and $n_{D\max}$ be the indices where $E_{A1}[n]$ and $E_{D1}[n]$ are maximum, respectively.
3. Find n_A such that $E_{A1}[n_A] < \mathrm{Th}_1$ and $n_{A\max} - n_A$ is minimum. Also find n_D such that $E_{D1}[n_D] < \mathrm{Th}_1$ and $n_{D\max} - n_D$ is minimum. The start syllable location n_B is assigned to $min(n_A, n_D)$.
4. Find n_A such that $E_{A1}[n_A] < \mathrm{Th}_1$ and $n_A - n_{A\max}$ is minimum. The stop syllable location is equal to n_A.
5. For vowel border determination, first find n_A such that $E_{A1}[n_A] < \mathrm{Th}_2$ and $n_{A\max} - n_A$ is minimum. And find n_D such that $E_{D1}[n_D] < \mathrm{Th}_2$ and $n_{D\max} - n_D$ is minimum. Then compute S_A, the sum of $E_{A1}[n]$ from $n = n_B$ to n_A and S_D, the sum of $E_{D1}[n]$ from $n = n_B$ to n_A. If $S_A < S_D$ then we select n_A as the start vowel location, otherwise n_D is the start vowel point. Next find n_D such that $E_{D1}[n_D] < \mathrm{Th}_2$ and $n_D - n_{D\max}$ is minimum. The stop vowel is equal to n_D as shown in Figure 4(e).

4. Experimental Results

To verify the proposed phoneme segmentation method, some experiments have been performed using the speech data of 1,000 syllables recorded from 10 speakers (5 males and 5 females, 21–26 years old.) Daubechie-4 wavelet coefficients are employed in the signal decomposition step. We let $\mathrm{Th}_1 = 0.05$ and $\mathrm{Th}_2 = 0.2$ to locate the syllable and vowel borders in all experiments. In order to verify the accuracy of the proposed method, the computed border locations of voice signals are compared with those obtained from a linguistic expert. The result is considered correct if the border location difference between the proposed method's and the expert's is less than 14 milliseconds. The overall results are summarized in Table 1. It shows that the proposed method yields highly accuracy rate for the starting positions of initial consonant, vowel, final consonant and silence.

5. Conclusions

In this paper, a new technique for Thai isolated word phoneme segmentation is proposed. The main feature of this method is the application of the energy contour of low and high frequency components of a speech signal. Signal decomposition is performed by simply applying wavelet decomposition. A heuristic algorithm for phoneme segmentation is described. The experimental results show that the proposed method can determine the border locations of the phoneme. It should be noted that, in the proposed method, the threshold values Th_1 and Th_2 are manually selected. The future plan to improve the method is to apply an adaptive technique to adjust these threshold values.

Table 1. Percentage of correct phoneme segmentation.

Speaker #	Ci	V	Cf	S
1	97	90.4	94.1	100
2	95.6	89.6	90.4	99.3
3	94.1	87.4	92.6	97.8
4	96.3	91.1	94.1	99.3
5	94.8	90.4	92.6	98.5
6	97	91.9	91.1	97.8
7	95.6	88.1	91.9	98.5
8	96.3	91.1	94.1	98.5
9	97	88.1	92.6	100
10	96.3	91.1	93.3	99.3
Average	96	89.9	92.7	98.9

Note: Ci, V, Cf, and S the starting positions of initial consonant, vowel, final consonant and silence, respectively.

References

[1] R. Gemello, D. Albesano, and F. Mana, "Multi-Source Neural Networks for Speech Recognition," *Intl. Joint Conf. on Neural Networks*, Vol. 5, pp. 2946-2949, 1999.

[2] L. Jun, X. Zhu, and Y. Luo, "An Approach to Smooth Fundamental Frequencies in Tone Recognition," *Proc. Communication Tech.*, Vol. 1, pp. S16-10-1 – S16-10-5, 1998.

[3] J. H. Lee and S. B. Rhee, "A study on consonants/vowels phonetic segmentation of Korean isolated words based on a rule-based system for the phenomenon of Korean vocalization," *Proceedings of IEEE Region 10 International Conference*, pp. 1351-1354, 1999.

[4] O. Maeran, V. Piuri, and G. Storti Gajani, "Speech recognition through phoneme segmentation and neural classification," *IEEE Instrumentation and Measurement Technology Conference*, pp. 1215-1220, 1997.

[5] F. R. McInnes, M. A. Jack, and J. Laver, "Template Adaptation in an Isolated Word Recognition System," *IEEE Proc. Communications, Speech and Vision*, pp. 119-126, 1989.

[6] A. D. Poularikas, *The Transforms and Applications Handbook*,Florida, CRC Press, 1996.

[7] L. R. Rabiner and B. H. Juang, *Fundamental of Speech Recognition*, Prentice Hall, Englewood Cliffs, N.J., 1993.

[8] L. R. Rabiner and S. E. Levinson, "*Isolated and Connected Word Recognition – Theory and Selected Applications*," IEEE Trans. on Comm. Vol. COM-29, No. 5, pp. 621-659, 1981.

[9] C. Suwanchewasiri, "Thai speech recognition for speaker dependent 500 words vocabulary based on phonemic distinctive features of isolated syllables and neural network," *The Fifth National Computer Science and Engineering Conference*, pp. 59-69, 2001.