How to cite this paper:

Paponput Sopon, Thongparn Suksamer, Jantima Polpinij, & Rapeeporn Chamchong. (2017). A framework of Thai text retrieval using speech in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 517-522). Sintok: School of Computing.

A FRAMEWORK OF THAI TEXT RETRIEVAL USING SPEECH

Paponput Sopon[1], Thongparn Suksamer[2], Jantima Polpinij[3], Rapeeporn Chamchong[4]

Intellect Laboratory, Faculty of Informatics, Mahasarakham University, Thailand
[1]*paphonput@gmail.com*, [2]*thongparn@gmail.com*,
[3]*jantima.p@msu.ac.th*, [3]*rapeeporn.c@gmail.com*

ABSTRACT. Today, speech-based information retrieval is necessary because of the various keyboard less applications such as user-friendly interface for PCs, car navigation systems, or mobile-phones. Unfortunately, current programs may not support for every language such as Thai language. Therefore, the goal of this work is to develop a system to retrieve documents by Thai speech. The proposed prototype application provides for desktop device such personal computer (PC). This application will accept Thai speech queries and then, it provides Thai text search result on the PC screen.

Keywords: Information Retrieval, Speech signal processing, Thai, Speech-based Information Retrieval

INTRODUCTION

The most IR systems have been driven on text-based information retrieval (Allan, 2002; Lee & Pan, 2009; Oliveira *et al*., 2011). This is because a large body of data is also available as complex unstructured data sources such text. Nowadays, text-based information retrieval is very successful because it has been studied for decades (Oliveira *et al*., 2011). However, speech-based information retrieval is required today because of the numerous keyboard-less applications such as user-friendly interface for personal computers (PCs), car navigation systems, or mobile-phones (Fujii *et al*. 2002; Anguera *et al*., 2014). To make easier for finding of relevant information on Web and the PCs, information retrieval by speech is required. It is necessary not only for any person in general but also for disabled persons such as handicapped persons or blind persons to access information on their PC (Oliveira *et al*., 2011).

Traditionally, an information retrieval (IR) system is to provide users with those documents that will satisfy their information need (Allan, 2002; Lee & Pan, 2009). IR allows easy access to huge amount of information (or data) (Lee & Pan, 2009). It includes the use of algorithms to process a huge of unstructured or semi-structured data. The most IR systems have been driven on text-based information retrieval (Allan, 2002; Lee & Pan, 2009). Today, text-based information retrieval is quite successful for IR systems because it has been studied and investigated for decades (Allan, 2002; Lee & Pan, 2009). Today, with the numerous keyboard-less applications, speech-based information retrieval have been studied because it allows the user to input a query as speech. As the result, let q be a speech query, while each document in the collection will be provided their features. Before retrieving relevant documents, the speech query must be transformed into some kind of content features such as key-

words. Finally, documents can be retrieved by directly comparing them with the query (Papka *et al*., 1998; Fujii *et al*., 2002).

Due to the advances in speech recognition technology, proper integration of IR and speech recognition has been considered and studied by many researchers (Itou *et al*., 2001; Fujii *et al*., 2002). In the previous study, text-based information retrieval has been investigated for decades but the study on speech-based text retrieval has just begun (Oliveira *et al*., 2011). Unlike the text-based information retrieval, text documents cannot be retrieved by directly comparing them with the speech query. It needs a process of speech-to-text in order to transform speech query to text before the process of text retrieval is performed.

As above, this work also presents a methodology of speech-based information retrieval in PC, where the proposed system will support for Thai language. The proposed system will accept Thai speech queries and then, it provides Thai text search result on the PC screen.

The remainder of the paper is organized as follows. The literature review is described in Section 2. The proposed methodology is presented in Section 3, while the experimental results are shown in Section 4. Finally, we summarize our research and discuss our future work directions in Section 5.

THE FRAMEWORK

The proposed framework consists of two main components. There are *Thai speech modelling* and *speed-based Thai text retrieval*. Each component can be described as follows.

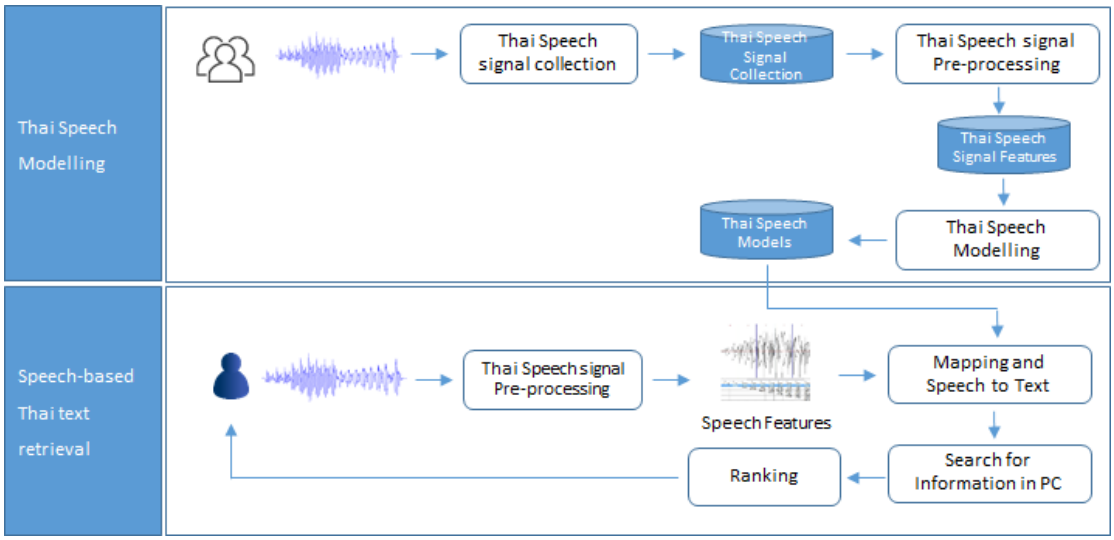

Figure 1. The Methodology of Speech-based Thai Text Retrieval

1. Thai speech modelling

It is to convert the speech signal formatted as audio wave (.wav) into a sequence of feature vectors. For Thai speech modelling in this work, it consists of three main processing steps. There are speech pre-processing, speech feature extraction, and acoustic modeling.

Speech pre-processing

It consists of three main steps such as pre-emphasis, endpoint detector, and windowing. *Pre-emphasis* is used to remove noise (Vergin & Shaughnessy, 1995) because noise can significantly decrease the performance of a speech recognition system (Vergin & Shaughnessy, 1995). Afterwards, it is to separate a speech signal into a short signal, called *endpoint detector* (Shen *et al*., 1998). To detect the start and end point of words in speech signal, the *cosine-based short-time energy* technique has been applied. Then, this technique is a modification of short-time energy, which is developed for this work. In the final stage of the speech pre-processing, the *windowing* is applied to separate each word speech into short windows of *N* samples, represented as *w*(*n*) by using the *Hamming* technique (Nuttall, 1981). In this work, a frame is in the size of 10-30 ms. range with possibly overlapping frames about *1/3* the size of the frame.

Speech feature extraction

After the speech pre-processing is done, it is the speech feature extraction. This work applies the Mel frequency cepstral coefficient (MFCC) algorithm to extract speech features because this algorithm is successfully employed to extract features from the speech signal. To obtain MFCC coefficient (Zheng *et al*., 2001), the input speech signal is windowed and taken *Discreet Fourier transform* (*DFT*) to convert into frequency domain by smoothing the *Fast Fourier Transform* (*FFT*) with around 20 frequency bins distributed non-linearly across the speech spectrum. The nonlinear frequency scale is called a *mel scale*[1]. Here, bank filter is used to wrapping the *mel scale*, and then a log magnitude of each of the *mel scale* is acquired.

Speech recognition process

It stage is to transform spoken words into text (Huang *et al*., 2014). In general, it is to automatically extract the string of words spoken from the speech signal. The recognized words can be used for applications such as information search, commands and control system, and data entry. In general, a process of speech recognition starts with the language model development. Speech signal will be firstly converted to a sequence of feature vectors based on spectral, and then the features of each word will be recognized by *Hidden Markov Models* (*HMM*) (Gales & Young, 2007). Finally, it will return the speech models that will be used to predict a set of words in speech signal input.

2. Thai text retrieval by speech

This is a mechanism of an information retrieval (IR) system that users will use to retrieve the relevant textual documents. Firstly, it accepts speech queries, and then the speech queries will be transformed to textual query through the use of the speech models. Simply speaking, the language models are used to predict a set of words contained in a speech query. It is a process of speech-to-text.

After having a textual query, this work applies the concept of keyword-search to find the relevant textual documents (Bao *et al*., 2010). This technique will look for words anywhere in the document. However, keyword-search often returns the results that include many non-relevant documents. Therefore, in IR, there is a need to rank documents by their relevance to the query. Relevance ranking (Clarke *et al*., 2000) is the method that is used to order the results list in such a way that the documents most likely to be of interest to a user will be at the

[1] A *Mel* is a unit of measure based on the human ear's perceived frequency.

front. This makes searching easier for users as they won't have to spend as much time looking through documents for the information that interests them. Finally, the system returns the results of textual search on the personal computer screen.

THE EXPERIMENTAL RESULTS

Dataset

There are two types of datasets used in this work. Firstly, it is a speech dataset. It will be used to train for acoustic models. These speeches are collected from 50 different persons, which can be divided into two groups - typically a female group and a male group. The dataset include 1,000 speech files, each speech signal containing a sequence of words up to 3-7 words.

The second dataset is a textual dataset relating to illness and medicine household in everyday life. A Thai text data is formatted as text files. This dataset consist of 500 documents. It will be used for experiment in Thai text retrieval by speech.

The Experimental Techniques

(1) The Experimental Results of Thai Speech Recognition Component

In general, word error rate (WER) (Klakow & Peters, 2002) is a common technique of the performance of a speech recognition system. WER can then be computed as:

$$WER = \frac{S + D + I}{N} \quad (1)$$

Consider the equation (1). Let *S* be the number of substitutions, *D* be the number of deletions, and *I* be the number of insertions. Also, *C* is the number of the correct, *N* is the number of words in the reference (*N* = *S* + *D* + C). Sometimes, *word accuracy* (*WA*) (Klakow & Peters, 2002) can be used instead:

$$WA = 1 - WER \quad (2)$$

This work experiments with 75 speech signals containing a sequence of words up to 5 words. The result of WA is 74.50% for the Thai speech recognition component. It can be seen that this component is good in terms of word accuracy. However, the result still shows a failure rate. Basically, speech recognition usually depends on "*distinctive features*" of speech signal in order to estimate the acoustic model. If given speech features are inappropriate, this can lead to lower recognition rates.

(*2*) *The Experimental Results of Thai Speech Recognition Component*

Common performance measures for system evaluation are *precision* (*P*), *recall* (*R*), and *F*-measure (Baeza-Yates & Ribeiro-Neto, 1999), and these techniques are also used in this works. After testing by the information retrieval measurement standard, the experimental results show the recall rate at 73%, while the precision rate is 71%. Finally, the result of *F*-measure is 71.50% for our proposed methodology.

In fact, the system of speech-based Thai text retrieval is can shows a failure rate. This is because it depends on the accuracy of the Thai speech recognition component. When it accepts speech queries, and then the speech queries will be transformed into textual queries. Therefore, if the accuracy of Thai speech recognition component is poor, the component of speech-based Thai text retrieval will return the failure rate as well.

CONCLUSION

This work aims to present a method of Thai text document retrieval using speech. This method will accepts a user's speeches as queries, and then it provides search results for display on a personal computer screen. The proposed method consists of two main components. They are speech modelling and speech-based text retrieval.

The speech modelling is the process of acoustic model creation. An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each acoustic model can be modelled by taking a large database of speech (called a speech corpus) and using special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models. For the speech-based text retrieval component. This component consists of two main processing steps. Firstly, it is the *speech recognition process*, where it is the process by which a computer maps an acoustic speech signal to text by using the acoustic models, and then this text will be used as queries. Next, it is the process of finding documents relevant to an information need from a large document set, called *Thai text retrieval process*.

After testing, the results of Thai speech recognition show an average word accuracy of 74.50%, while the result of speech-based Thai text retrieval 71.50% after testing by *F*-measure.

REFERENCES

Anguera, X., Rodriguez-Fuentes, L., Szöke, I., Buzo, A., & Metze, F. (2014). Query by Example Search on Speech at Mediaeval. *Proceedings of the MediaEval 2014 Workshop*.

Allan, J. (2002). Perspectives on Information Retrieval and Speech. *Lecture Notes in Computer Science: Vol. 2273. Information Retrieval Techniques for Speech Applications (pp. 1-10). Berlin, Germany: Springer-Verlag.* doi: 10.1007/3-540-45637-6_1

Bao, Z., Lu, J., Ling, T.W., & Chen, B. (2010). Towards an Effective XML Keyword Search, *IEEE Transactions on Knowledge and Data Engineering*, 22(8), 1077-1092. Doi: 10.1109/ TKDE. 2010.63

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York. NY: ACM Press.

Clarke, C., Cormackb, C.A., Tudhope, E.A. (2000). Relevance ranking for one to three term queries. *Information Processing and Management*, 36(2), 291-311.

Fujii, A., Ituo, K., & Ishikawa, T. (2002). A Method for Open-Vocabulary Speech-Driven Text Retrieval. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 10, 188-195. doi: 10.3115/1118693.1118718

Gales, M. & Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. *Journal Foundations and Trends in Signal Processing*, 1(3), 195–304. doi: 10.1561/2000000004

Huang, X., Baker, J., & Reddy, J. (2014). A historical perspective of speech recognition. *Magazine Communications of the ACM*, 57(1), 94-103. doi: 10.1145/2500887

Itou, K., Fujii, A., & Ishikawa, T. (2002). Language modeling for multi-domain speech driven text retrieval. *IEEE Automatic Speech Recognition and Understanding Workshop*. doi: 10.1109/ASRU.2001.1034653

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2), 19–28. doi: 10.1016/S0167-6393(01)00041-3

Lee L.S., & Pan, Y.C. (2009). Voice-based information retrieval — how far are we from the text-based information retrieval? *IEEE Workshop on Automatic Speech Recognition & Understanding*, 26-43. doi: 10.1109/ASRU.2009.5372952

Nuttall, A.H. (1981). Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, Signal Processing*. 29(1), 84–91. doi: 10.1109/TASSP.1981.1163506

Oliveira, J, Guerreiro, T, Nicolau, H., Jorge, J., & Gonçalves, D. (2011). Blind People and Mobile Touch-based Text-Entry: Acknowledging the Need for Different Flavors. *The proceedings of the*

13th international ACM SIGACCESS conference on Computers and accessibility, 179-186. doi: 10.1145/2049536.2049569

Papka, R., Callan, J.P., & Barto, A.G. (1996). Text-based Information Retrieval using Exponentiated Gradient Descent. *Proceedings of the 10th Annual Conference of Advances in Neural Information Processing Systems*, 3-9.

Rahmouni, R.H, & Sayadi, M. (2004). Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier. *Proceedings of IEEE First International Symposium on Control, Communications and Signal Processing*, 631-634. doi: 10.1109/ISCCSP.2004.1296479

Shen, J., Hung, J., & Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments, *The 5th International Conference on Spoken Language Processing*.

Vergin, R., & O'Shaughnessy, D. (1995). Pre-emphasis and speech recognition. *Canadian Conference on Electrical and Computer Engineering*, 1062-1065. doi: 10.1109/CCECE.1995.526613

Zheng, F., Zhang, G., & Song, Z. (2001), Comparison of Different Implementations of MFCC. *Journal of Computer Science and Technology*, 16(6): 582–589